Instituto Nacional de Estadística

Oposiciones al Cuerpo de Diplomados en Estadística del Estado

BOE núm. 270, de 12 de octubre de 2020, pág. 87165

Estadística

INE
02/07/2021

Índice general

Tema 24

Introducción a la depuración e imputación de datos estadísticos en el proceso estadístico. Datos, errores, datos ausentes y controles (edits). Métodos básicos para la depuración e imputación de datos estadísticos. Estrategia de depuración e imputación.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

T. de Waal, J. Pannekoek y S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Amsterdam: Wiley

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

24.1 Introducción a la depuración e imputación de datos estadísticos en el proceso estadístico

El objetivo de los Institutos Nacionales de Estadística (INEs) es proporcionar estadísticas de gran calidad sobre muchos aspectos de la sociedad, tan actualizadas y exactas como sea posible. Una de las dificultades que surgen a lo largo del proceso de obtención de las estadísticas es el hecho de que tanto las encuestas tradicionales como los datos administrativos que se usan contienen errores que pueden influir en las estimaciones. Con el fin de evitar sesgos e inconsistencias en la publicación de los datos, el INE realiza un proceso de chequear los datos recogidos y corregirlos en caso de que sea necesario. Este proceso de mejora de la calidad de los datos mediante la detección y corrección de errores comprende una gran variedad de procesos, tanto manuales como automáticos, que se denominan *depuración de datos estadísticos*. La depuración de datos estadísticos se lleva estudiando desde mediados de los años 50 (véase, p.ej., Nordbotten 1955).

Además de errores en los datos, otro factor que complica el trabajo de los INEs es la existencia de datos *missing* (datos ausentes). Esto se puede considerar como otra forma de datos erróneos, que son fáciles de identificar, pero para los que resulta difícil estimar un buen valor.

Los errores aparecen durante el proceso de medida cuando los valores proporcionados difieren de los valores verdaderos. Esto se puede deber a que los verdaderos valores son desconocidos, difíciles de conseguir. Otro motivo podría ser que las preguntas son mal interpretadas o mal leídas por los informantes. Un ejemplo es el denominado error de unidad de medida que ocurre si el informante proporciona los datos en euros cuando se le pide que los indique en miles de euros. Otro ejemplo es que el informante proporcione sus propios ingresos cuando se piden los ingresos del hogar y el hogar está compuesto por más personas además del informante. En el caso de encuestas económicas, los errores también tienen lugar debido a que las definiciones usadas por los INEs no coinciden con las usadas por el sistema contable de la unidad informante. Puede haber, por ejemplo, diferencias en el periodo de referencia usado por las empresas y el periodo solicitado (el año fiscal frente al año natural es un ejemplo). Después de que los datos han sido recogidos, pasarán por varios procesos, como la codificación, la depuración y la imputación. Los errores que surgen a lo largo de estos otros procesos se conocen como errores de procesamiento. Señalamos que, aunque el propósito de la depuración es corregir los errores, también debe tenerse presente que, como proceso, la depuración también puede introducir errores de forma ocasional. Esta situación no deseable surge si el valor de una variable se modifica porque parece ser erróneo cuando en la realidad es correcto. Los datos *missing* aparecen cuando un informante no sabe la respuesta a una pregunta o se niega a dar la respuesta a una determinada pregunta.

Tradicionalmente, los INEs siempre se han esforzado e invertido muchos recursos en la depuración de los datos, ya que se considera un requisito muy importante para publicar estimaciones acuradas y con calidad. En los procesos tradicionales de procesamiento de una encuesta, la depuración era principalmente interactiva (manual) con la intención de corregir todos los datos en detalle. Los errores detectados o las inconsistencias eran corregidas después de contactar con el informante, lo que implicaba un trabajo que requiere mucho tiempo y trabajo. En este tema se verán métodos más eficientes de depuración.

Se ha admitido desde hace tiempo que no es necesario corregir todos los datos en detalle. Varios estudios (véanse, p.ej., Granquist 1984; Granquist 1995; Granquist 1997b; Granquist 1997a; Granquist 1997c; Granquist y J.G. Kovar 1997) han mostrado que, en general, no es necesario eliminar todos los errores de un conjunto de datos para obtener valores agregados publicables fiables. Los principales productos estadísticos son tablas que contienen agregados, que a menudo se basan en muestras de la población. Esto implica que pequeños errores en registros individuales son aceptables. En primer lugar, porque estos errores tienden a cancelarse cuando se agregan. En segundo lugar, porque si los datos se obtienen a partir de una muestra, siempre habrá un error de muestreo en los valores publicados, incluso si los datos recogidos son completamente correctos. Con el fin de obtener datos de suficiente calidad, normalmente es suficiente con eliminar sólo los errores más influyentes.

Se emplea demasiado esfuerzo corrigiendo errores que no tienen un impacto notable en los valores publicados. Esto se denomina 'sobredepuración'. La sobredepuración no

sólo implica un coste, sino que conlleva una gran cantidad de tiempo, que hace que el periodo entre la recogida de datos y la publicación sea innecesariamente largo. En ocasiones, la sobredepuración se llega a convertir en una 'depuración creativa', que incluso resulta negativa para la calidad de los datos, ya que se modifican datos que son correctos. Para más información sobre los riesgos de la sobredepuración y la depuración creativa[1] véanse (Granquist 1995; Granquist 1997c; Granquist y J.G. Kovar 1997).

Se ha sostenido que el papel de los INEs en la depuración no debería reducirse a la detección y corrección de errores. Granquist (1995) identifica los siguientes objetivos:

1. Identificar la fuente de errores para proporcionar *feedback* sobre el proceso de producción completo.
2. Proporcionar información sobre la calidad de los datos iniciales y finales [2].
3. Identificar y tratar los errores influyentes y los *outliers* en los datos individuales.
4. Cuando sea necesario, proporcionar datos individuales completos y consistentes.

Los datos *missing* constituyen un problema bien conocido al que tienen que enfrentarse todos los organismos que recojan datos sobre personas o empresas. Dependiendo de la legislación existente puede ser más o menos importante en cada país. La solución más común es la imputación, donde los valores de los datos *missing* son estimados. Un problema importante de la imputación es preservar la distribución estadística del conjunto de datos. Éste no es un problema sencillo, especialmente para datos de grandes dimensiones.

En los INEs el problema de la imputación es aún más complicado debido a la existencia de limitaciones en forma de restricciones en los edits, o edits a secas, que los datos tienen que satisfacer. Ejemplos de tales edits son que los beneficios y los costes de una empresa tienen que sumar su cifra de negocios. Los registros que no satisfagan este edit son inconsistentes y por tanto se consideran incorrectos.

24.2 Datos, errores, datos ausentes y controles (edits)

Durante el proceso de depuración y de imputación de los datos, los registros erróneos, y los valores erróneos dentro de estos registros, se localizan y se estiman nuevos valores para los valores erróneos y los valores *missing*. La depuración consiste en llevar a cabo dos pasos: primero se localizan los valores incorrectos, a esto se le llama a menudo *localización del error*, y a continuación los valores tienen que ser *imputados*, es decir, se tienen que sustituir por valores mejores, preferiblemente, los correctos.

[1] Para las encuestas con estimaciones basadas en muestreo probabilístico, es fácil aprehender el riesgo de manipulaciones excesivas de los datos, pues se están introduciendo fuentes de variabilidad que no se controlan, introduciendo, por tanto, sesgos desconocidos y aumentando la varianza real de las estimaciones.

[2] Se denominan datos iniciales a la primera versión proporcionada por los informantes y datos finales a los datos una vez depurados y validados

En principio no es necesario imputar los datos *missing* ni los valores erróneos para obtener estimaciones válidas. En su lugar, se pueden estimar las variables objetivo directamente durante la fase de estimación, sin imputar los datos *missing* ni los erróneos. Sin embargo, este enfoque es en la mayoría de los casos prácticos muy complejo. Mediante una primera imputación de los valores *missing* y los erróneos, se obtiene un conjunto completo de datos. Y a partir de este conjunto completo de datos, se obtienen las estimaciones mediante métodos de estimación estándar. Por tanto, la imputación a menudo se lleva a cabo para simplificar el proceso de estimación.

Las técnicas de depuración e imputación se pueden dividir en dos clases principales, dependiendo del tipo de datos a depurar o imputar: técnicas para datos numéricos y técnicas para datos categóricos (datos entre los cuales no hay una relación de orden, datos agrupados o datos para variables ficticias -*dummy*-). Generalmente, hay diferencias importantes entre las técnicas para estos tipos de datos. Los datos numéricos sobre todo se recogen en encuestas económicas (empresas, establecimientos) mientras que los datos categóricos se recogen en encuestas sociales (personas, hogares, viviendas).

La depuración de encuestas económicas suele ser un problema más complejo que la de la mayoría de las encuestas sociales. Dentro de las encuestas económicas distinguimos entre las encuestas coyunturales, pocas variables y con mucha periodicidad (mensual o trimestral) y las encuestas estructurales, con muchas variables, muchas desagregaciones y periodicidad anual. La principal razón es que en las encuestas económicas estructurales hay muchas más reglas de depuración que en las encuestas sociales y las encuestas económicas estructurales contienen muchos más errores que las sociales.

En los últimos años se ha incrementado el uso de datos administrativos en los INEs. La depuración e imputación de datos administrativos para fines estadísticos tiene determinadas características que no se encuentran en las encuestas muestrales. Por ejemplo, si los datos de varios registros se combinan, además de los errores presentes en los registros individuales, también podemos encontrarnos inconsistencias adicionales entre los datos de los distintos registros debidos a los errores que se producen al cruzar los registros o las divergencias debidas a las definiciones en los metadatos. Véase (A. Wallgren 2007) para una descripción sobre los métodos para las estadísticas basadas en registros administrativos.

24.2.1 Tipos de errores

Uno de los objetivos más importantes de la depuración de datos es la detección y corrección de errores. Los errores se pueden clasificar de varias formas.
Una primera distinción importante se hará entre errores sistemáticos y aleatorios. La segunda será entre errores influyentes y no influyentes. La última será entre *outliers* y no *outliers*.

Definición 1

Errores sistemáticos.
Este tipo de errores puede ocurrir cuando un informante malinterpreta o lee incorrectamente una pregunta. Los errores sistemáticos pueden dar lugar a agregados sesgados. Una vez que se detectan, los errores sistemáticos se pueden corregir fácilmente porque se conoce el mecanismo subyacente. Este mecanismo se puede observar bien a lo largo de toda la historia de un informante o transversalmente a la muestra.

Un error sistemático bien conocido es el llamado error de unidad de medida. Este error ocurre cuando un informante proporciona el valor de una variable en una unidad de medida errónea. Por ejemplo, supongamos que la cifra de negocios total tiene que ser proporcionada en miles de euros, pero se declara en euros.

Los errores sistemáticos, como los errores en las unidades de medida, se pueden detectar a menudo comparando el valor actual de un informante con el de períodos anteriores (meses, años, trimestres), comparando las respuestas a las variables del cuestionario con los valores de variables de registros, o usando el conocimiento de un experto. Los errores de redondeo se pueden detectar probando si los edits de balance que no se verifican lo hacen con un pequeño cambio en el valor de las variables afectadas.

Otra forma de detectar los errores sistemáticos es usando edits de razón, que fija unos límites aceptables para un cociente de dos variables. Por ejemplo, considerando el ejemplo anterior del error en la cifra de negocios, si se dispone de la cifra de empleados y esta variable no se ve afectada por el mismo error, el edit basado en el cociente entre la cifra de negocios y los empleados serviría para detectar a unidades que incurran en este error (ya que los valores de los cocientes serían anómalos comparados con los valores de los no erróneos).

Una posible causa de errores sistemáticos que hay que intentar evitar es la causada por un encuestador que ha entendido mal la información a recoger. Por este motivo es muy importante la formación del personal de recogida de datos y que las dudas que se planteen durante la recogida sean transmitidas a los expertos en la materia.

Otros ejemplos de errores sistemáticos son los siguientes: errores asociados con un malentendido sobre los filtros de las preguntas en el cuestionario; errores debidos a la falta de habilidad del informantes para proporcionar información basada en una clasificación/definición estadística específica; errores en signos que pueden ocurrir cuando un informante omite de forma sistemática un signo negativo en alguna variable, por ejemplo los beneficios, que puede ser negativa; valores *missing* sistemáticos como no indicar el valor total para algunas variables.

Errores aleatorios.
Los errores aleatorios son debidos al azar, son accidentales. Un ejemplo es un valor observado donde un informante por error tecleó un dígito de más. En la estadística, en general, la esperanza de un error aleatorio es cero. Sin embargo, en nuestro caso, la esperanza de un error aleatorio puede no ser cero. Este es, por ejemplo, el caso del ejemplo anterior.

Los errores aleatorios pueden dar lugar a valores atípicos. En tal caso se pueden detectar usando técnicas de detección de *outliers* o de depuración selectiva. Los errores aleatorios también pueden ser influyentes, en cuyo caso pueden ser detectados con técnicas de depuración selectiva. Si los errores aleatorios no dan lugar a valores atípicos o a errores influyentes se pueden corregir de forma automática.

Si el error aleatorio no es un valor atípico ni influyente la forma de detectarlo es porque no se verifican alguna regla de depuración. Por ejemplo, si en un cuestionario se indica que la edad es '12' y en estado civil figura 'Casado'.

Hay varios principios básicos en la localización de campos erróneos en un registro con inconsistencias. Los principios más conocidos y más utilizados es el paradigma de Fellegi y Holt 1976.

Errores influyentes.
Los errores que tienen una gran influencia en los valores publicables se llaman errores influyentes. Pueden ser detectados con técnicas de depuración selectiva.

El hecho de que un valor tenga una gran influencia en las estimaciones no implica necesariamente que el valor sea erróneo. De hecho, en las encuestas económicas las observaciones influyentes son bastante comunes por dos motivos. Por un lado, un pequeño número de unidades (empresas o establecimientos) pueden ser mucho más grandes que otras, en términos del número de empleados o de la cifra de negocios. Por otro lado, algunas unidades pueden tener pesos de muestreo muy grandes, incluso aunque esas unidades no sean grandes, su contribución a las estimaciones puede ser significativa.

Outliers.
Un valor, o un cuestionario, se denomina *outlier* si no se ajusta bien a un modelo considerado para los datos observados. Si un único valor es un *outlier*, se llama *outlier* de una variable. Si el cuestionario en su totalidad, o al menos un subconjunto de varios valores, es un *outlier* cuando los valores se consideran de manera simultánea, se denomina *outlier* multivariante. De nuevo, el simple hecho de que un valor sea un *outlier* no implica necesariamente que este valor contenga un error.

El modelo que se utiliza para considerar una observación como *outlier* se refiere a una población base, no al total de la muestra, y a menudo hay distintos modelos para las distintas subpoblaciones. Por ejemplo un modelo puede ser apropiado sólo para empresas con una actividad económica particular. La división 30 de la CNAE ('Fabricación de otro material de transporte') incluye empresas de construcción naval, aeronáutica y de material ferroviario cuyas cifras de negocios se pueden considerar *outliers* pero que no lo son.

Los *outliers* están relacionados con los valores influyentes. Un valor influyente a menudo es también un *outlier*, y viceversa. Sin embargo, un *outlier* también puede ser un valor no influyente y un valor influyente también puede no ser un *outlier*. Los *outliers* a menudo se detectan durante la macrodepuración.

24.2.2 Tipos de datos *missing*

Los datos *missing* implican una reducción del tamaño efectivo de muestra (que se puede resolver sobremuestreando), y en consecuencia un incremento del error cometido en la estimación (que se debe cuantificar mediante la estimación de dicho error). Un efecto más problemático, que no se puede medir fácilmente, es el sesgo de las estimaciones. Si el mecanismo en la falta de respuesta no depende de datos no observados, la imputación puede dar lugar a estimaciones insesgadas sin la necesidad de hacer ninguna hipótesis. En el caso contrario es necesario hacer nuevas hipótesis para reducir el sesgo mediante la imputación.

Una clasificación de los mecanismos de falta de respuesta que se usa a menudo es: completamente *missing* aleatoriamente (MCAR del inglés *missing completely at random*), *missing* aleatoriamente (MAR del inglés *missing at random*) y *missing* no aleatoriamente (NMAR del inglés *not missing at random*); véanse (Rubin 1987; Schafer 1997; R. Little y Rubin 2002).

Definición 2

MCAR.
La probabilidad de que un valor sea *missing* no depende del(los) valor(es) de la(s) variable(s) objetivo que se imputarán ni de los valores de las variables auxiliares: olvido de la respuesta o pérdida de parte de los datos durante su procesamiento. En este caso los datos observados se pueden considerar como un subconjunto aleatorio de los datos completos. Desgraciadamente, el MCAR raramente ocurre en la práctica. Formalmente:

$$\mathbb{P}(r_j|y_j, \mathbf{x}, \xi) = \mathbb{P}(r_j|\xi). \tag{24.1}$$

donde r_j es el indicador de respuesta de la variable objetivo y_j, donde $r_{ij} = 1$ si el registro i contiene una respuesta para la variable y_j, y $r_{ij} = 0$ en caso contrario, $\mathbf{x}$ es

un vector de variables auxiliares que siempre tendremos y ξ es un parámetro del mecanismo de falta de respuesta.

MAR.
La probabilidad de que un valor sea *missing* depende de un valor de las variables auxiliares, pero no depende del(los) valor(es) de la(s) variable(s) objetivo que se imputarán. Por ejemplo, el mecanismo de falta de respuesta para mayores es distinto del de los jóvenes, pero dentro de cada grupo no depende del valor de la variable objetivo; o en caso de encuestas económicas, las diferencias que se dan entre empresas grandes y pequeñas. Formalmente:

$$\mathbb{P}(r_j|y_j, \mathbf{x}, \xi) = \mathbb{P}(r_j|\mathbf{x}, \xi). \tag{24.2}$$

En este caso es necesario encontrar los grupos de unidades poblacionales adecuados para pasar del MAR al MCAR dentro de cada grupo.

NMAR.
La probabilidad de que un valor sea *missing* depende del(los) valor(es) de la(s) variable(s) objetivo que se imputarán y de los valores de las variables auxiliares: pregunta sobre ingresos, habrá más falta de respuesta en caso de altos ingresos. Formalmente:

$$\mathbb{P}(r_j|y_j, \mathbf{x}, \xi) \neq \mathbb{P}(r_j|\xi), \mathbb{P}(r_j|y_j, \mathbf{x}, \xi) \neq \mathbb{P}(r_j|\mathbf{x}, \xi). \tag{24.3}$$

Es el caso más complicado y no se puede usar únicamente los datos observados, sino que hace falta modelizar la dependencia de los mecanismos de falta de respuesta sobre el(los) valor(es) de la(s) variable(s) objetivo.

Otra clasificación de los mecanismos de falta de respuesta relacionados es:

Definición 3

Ignorable.
En caso de que sea MAR (o MCAR) y los parámetros a estimar sean "distintos"del parámetro ξ, es decir, el conocimiento de ξ no ayuda en la estimación de los parámetros de interés.

No ignorable.
Si el mecanismo es NMAR o el parámetro ξ no es "distinto"de los parámetros de interés o se dan ambos casos.

24.2.3 Reglas de depuración

Los errores en general se detectan con reglas de depuración o edits. Los edits definen los valores admisibles (o razonables) y las combinaciones de valores de las variables en cada cuestionario. Los errores se detectan verificando si los valores son admisibles de acuerdo con los edits, es decir, comprobando si los edits se verifican o no. Un edit se puede formular como

$$e : x \in S_x,$$

siendo S_x el conjunto de valores admisibles de x. Como veremos a continuación, x se puede referir a una única variable o a varias. Si e es falso, el edit no se cumple mientras que de lo contrario el edit se satisface.

Los edits de pueden clasificar en $duros$ o $blandos$. Los edits duros son aquéllos que se deben satisfacer para que un cuestionario sea considerado válido. Por ejemplo, un edit duro para una encuesta a una empresa especifica que la variable *Gastos totales* tiene que ser igual a la suma de las variables *Gastos de personal*, *Gastos de capital*, *Gastos de transporte*, y *Otros gastos*. Los cuestionarios en que no se verifiquen uno o más edits duros son considerados como inconsistentes y se deduce que alguna(s) variable(s) en el mismo es errónea. Los edits blandos se usan para identificar valores dudosos que se sospecha que pueden ser erróneos.

Algunos ejemplos son (a) un edit especifica que los salarios anuales de los empleados deben de ser inferiores a 10 millones de euros o (b) un edit especifica que la cifra de negocios por empleado de una empresa no puede ser mayor que 10 veces el valor del año anterior. Si no se verifica algún edit blando hay que seguir analizando los datos para confirmarlos o rechazarlos.

Cabe señalar que los edits se deben basar en el conocimiento sobre el tema, es decir, sobre el cocimiento de las condiciones sociales y económicas que pueden influir a los informantes y las implicaciones que tienen en la relación entre los apartados del cuestionario. Además, la definición de las regiones de aceptación de los distintos edits debería de estar respaldado por métodos estadísticos específicos. En particular, el análisis de las distribuciones conjuntas puede facilitar en gran medida la especificación de edits adecuados al mostrar relaciones entre las variables (Whitridge y J. Kovar 1990). Los métodos gráficos también pueden resultar útiles.

Veamos a continuación varios ejemplos de clases de edits.

Definición 4

Edits univariantes o restricciones de rango.
Un edit que describa los valores admisibles de una única variable se llama edit univariante o restricción de rango. Para variables categóricas una restricción de

rango simplemente verifica si los códigos de categoría observados para la variable pertenecen al conjunto especificado de código. El conjunto de valores permitidos S_x es

$$S_x = \{x_1, x_2, ..., x_C\},$$

y consiste en la enumeración de los C códigos permitidos. Por ejemplo, para la variable $Sexo$ podemos tener $S_x = \{0, 1\}$. Las restricciones de rango para variables continuas se especifican generalmente usando desigualdades. Las más sencillas son las restricciones de valores no negativos, es decir,

$$S_x = \{x | x \geq 0\},$$

Algunos ejemplos son $Edad$, *distintos tipos de costes*, etc. También son comunes restricciones de rango que describen un intervalo como

$$S_x = \{x | i \leq x \leq s\},$$

siendo i el límite inferior y s el superior. Algunos ejemplos son valores admisibles de edad, ingresos u horas trabajadas por semana.

Edits bivariantes.
En este caso el conjunto de valores admisibles de una variable x depende del valor de otra variable, que denominaremos y, observada en la misma unidad. El conjunto de valores admisibles es entonces el conjunto de pares admisibles de valores (x, y). Por ejemplo, si x es *Estado Civil* con valores 0 (nunca casado), 1 (casado) y 2 (previamente casado) e y es $Edad$, podemos tener

$$S_{xy} = \{(x, y) | x = 0 \wedge y < 16\} \cup \{(x, y) | y \geq 16\},$$

equivalente a $S_{xy} = \{(x, y) | x - y > 15\}$.

También podemos encontrarnos con edits de razón que se pueden definir como

$$S_x = \{(x, y) | i \leq \frac{x}{y} \leq s\},$$

Por ejemplo, el cociente entre la cifra de negocios y el número de empleados de una empresa en una determinada rama de la industria.

Edits de balance.
Los edits de balance son edits multivariantes que establecen que los valores admisibles de un número de variables están relacionadas con una ecuación lineal. Dos ejemplos son:

$$\begin{aligned} \text{Beneficios} &= \text{Cifra de negocios} - \text{Costes totales} \\ \text{Costes totales} &= \text{Gastos de personal} + \text{Otros costes} \end{aligned} \tag{24.4}$$

Los edits de balance son de gran importancia en la depuración de las encuestas económicas.

Como los edits de balance describen relaciones entre muchas variables se consideran edits multivariantes y deberían tratarse como un sistema de ecuaciones lineales. Es conveniente expresar dicho sistema con notación matricial. Si denotamos las variables de las restricciones 24.4 por x_1 (Beneficios), x_2 (Cifra de negocios), x_3 (Costes totales), x_4 (Gastos de personal) y x_5 (Otros costes), el sistema se puede escribir como

$$\begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

o como **Ax=0**. Los valores admisibles de un vector **x** sujeto al sistema de edits de balance, definido por la matriz de restricciones **A**, se puede escribir como

$$S_{\mathbf{x}} = \{\mathbf{x} | \mathbf{Ax=0}\}. \tag{24.5}$$

24.3 Métodos básicos para la depuración e imputación de datos estadísticos

Antes de explicar los métodos veamos por qué se han desarrollado estos métodos. Los ordenadores se han usado en el proceso de depuración desde hace muchos años, de hecho, en 1963 ya se mencionan en los *papers* sobre este tema. En los primeros años, sin embargo, su papel se limitaba a comprobar qué *edits* no se verificaban. Se grababan los datos en una base de datos, el ordenador comprobaba si los datos verificaban los *edits* especificados y para cada registro se listaban todos los *edits* que no se verificaban para corregir estos datos. Es decir, se subsanaban todos los cuestionarios en papel que no verificaban todos los *edits*. Este proceso iterativo continuaba hasta que (casi) todos los registros verificaban todos los *edits*.

El principal problema de este enfoque es que durante el proceso de depuración manual no se verificaba la consistencia de los registros. El resultado es que un registro que estaba 'correcto' podía incumplir uno o más *edits* especificados. Dicho cuestionario, por consiguiente, precisaba más corrección. No era excepcional que algunos registros tuvieran que ser corregidos varias veces. Por tanto, no es de extrañar que depurar de esta forma fuera muy costoso, tanto en términos de dinero como de tiempo, estimándose que entre un 25 % y un 40 % del presupuesto total se empleaba en la depuración.

24.3.1 Depuración durante la fase de recogida de datos

La técnica de depuración más eficiente de todas es no depurar, sino asegurarse de que los datos que se obtienen durante la fase de recogida son los correctos. Si el objetivo es recoger los datos correctos durante la recogida de los mismos, normalmente se usa un ordenador para grabar los datos. Cuando se da una respuesta inválida a una pregunta o existe una inconsistencia entre dos o más respuestas y la recogida se realiza usando un método asistido por ordenador (CAPI, CATI, CASI o CAWI) los errores pueden ser notificados de manera inmediata. (Para más información sobre datos recogidos por ordenador véase p.ej. Couper y col. 1998). De esta forma estos errores se pueden solucionar preguntando a los informantes de nuevo. Para CASI y CAWI normalmente no se programan todos los edits, ya que el informante se puede sentir molesto y puede negarse a completar el cuestionario cuando los edits saltan a medida que responde el cuestionario indicando que sus respuestas son inconsistentes.

Una vez terminada la fase de recogida por CAPI, CATI, CASI o CAWI estos cuestionarios contienen menos errores que lo recogidos mediante cuestionarios en papel ya que los errores aleatorios que afectan a los cuestionarios en papel no pueden ser detectados y corregidos durante la recogida. Además, si se recogen por CASI o CAWI se pueden evitar los edits de balance calculando de manera automática los totales a partir de las partes. Aunque hay evidencias de que los informantes pueden ser menos acurados cuando rellenan un cuestionario electrónico si los totales se calculan de manera automática.

Los INEs en los últimos años se han movido hacia el uso de recogida de datos usando *mixed-modes* donde los datos se recogen usando una mezcla de varios métodos de recogida de datos (véase p.ej. Leeuw 2005). Esto, obviamente, tiene consecuencias para la depuración de datos.

Un primer inconveniente es que puede resultar costoso a corto plazo, pero no a largo plazo. Otro inconveniente es que los informantes tienen que ser capaces de responder durante la entrevista, lo que en el caso de las encuestas económicas no es sencillo. En el caso de CAWI esto se puede solucionar fácilmente si se permite responder a la encuesta en varias etapas.

24.3.2 Métodos modernos de depuración

Depuración interactiva.
El conocimiento de los expertos se debe utilizar en la medida de lo posible desarrollando herramientas de depuración interactiva efectivas que permitan comprobar los edits específicos durante la recogida o una vez terminada, y, en caso de que sea necesario, corregir los datos erróneos de manera inmediata. Esto es lo que se denomina depuración interactiva o asistida por ordenador.

Para corregir los datos erróneos se pueden seguir varios métodos: el informante puede ser contactado de nuevo, los datos se pueden comparar con los de años/meses previos, los datos se pueden comparar con datos de informantes similares, o se puede usar el conocimiento del experto. Hoy en día es un método estándar de depuración tanto para datos numéricos como para categóricos. El número de variables, edits y registros puede ser, en principio, alto. Y la calidad de los datos depurados de esta forma se considera alta.

Depuración selectiva.
La depuración selectiva es un término general para varios métodos de detección de errores influyentes y *outliers*. Las técnicas de depuración selectiva tienen por objetivo aplicar depuración interactiva a un subconjunto de registros bien elegidos de forma que el tiempo y los recursos limitados disponibles para la depuración interactiva se empleen en esos registros que afectan más a la calidad de las estimaciones finales a publicar. Las técnicas de depuración selectiva intentan conseguir este objetivo dividiendo los datos en dos flujos: los registros del flujo crítico se depuran de manera tradicional interactiva, mientras que los registros no críticos se depurarán de forma automática.

Macrodepuración.
Distinguiremos entre dos formas de macrodepuración. La primera forma se llama a veces el método de agregación. Formaliza y sistematiza lo que todos los INEs hacen antes de la publicación: verificar si las cifras que se publicarán parecen razonables. Esto se lleva a cabo comparando las cantidades de las tablas a publicar con las mismas cantidades en publicaciones anteriores o con publicaciones relacionadas. Sólo en el caso de que se observe un valor inusual, se usa un proceso de microdepuración a los registros individuales y a los campos que contribuyen a la cantidad sospechosa.

Una segunda forma de macrodepuración es el método de la distribución. Los datos disponibles se usan para caracterizar la distribución de las variables. A continuación, todos los valores individuales se comparan con la distribución. Los registros que contengan valores que se puedan consideran extraños, teniendo en cuenta la distribución, son candidatos para una mayor inspección y posiblemente para depuración. La macrodepuración, en particular, el método de agregación, siempre se ha usado en los INEs.

Depuración automática.
Cuando la depuración automática se utiliza, los registros son depurados por un ordenador sin la intervención humana. En este sentido, la depuración automática es lo contrario a la aproximación tradicional en el problema de depuración, donde cada registro se depura manualmente. En los últimos años esta depuración se ha perfeccionado mucho ya que los ordenadores son más rápidos y los algoritmos se han simplificado y se han vuelto más eficientes.

24.3.3 Métodos de imputación

La imputación consiste en asignar un valor a un ítem o un grupo de ítems que previamente no tenia valor o ese valor se consideraba erróneo o no ajustado a la realidad. La

imputación es por lo tanto un proceso por el que se generan valores artificiales y por lo tanto introduce un error de imputación. Sin embargo, este error cuenta con la ventaja de ser medible ya que el especialista puede analizar la precisión de las imputaciones y de esta forma estimar el error de imputación.

La metodología para las estadísticas económicas modernas (Eurostat 2014) distingue tres grandes enfoques para la imputación. El primero consiste en la imputación deductiva o lógica, que consiste en usar reglas de derivación con la información disponible para estimar el valor *missing*. El segundo consiste en usar reglas de predicción estadísticas para obtener modelos donde calcular imputaciones. El tercer grupo consiste en utilizar unidades similares para imputar con este valores a las unidades donde no hay respuesta. Además de estos tres existe también en enfoque más manual de la imputación que consiste en la imputación por el experto en la materia. Veamos los principales métodos.

Imputación deductiva.
Este es el método que tiene preferencia sobre todos los demás pero en muchos de los casos no puede ser usado. Se trata de un método especialmente interesante cuando tenemos falta de respuesta parcial ya que utilizando el valor de otros ítems se puede deducir el valor *missing*. Por ejemplo, teniendo la cifra de negocios de España y del extranjero y no teniendo el total podríamos imputar el valor total como la suma de ambos.

Imputación basada en modelos.
Este enfoque de imputación consiste en encontrar el modelo predictivo adecuado para la obtención de la imputación. El modelo toma por lo tanto la información disponible y puede ser más o menos complejo. Dentro de estas técnicas destacan:

- **Imputación por la media**. Como su nombre indica cada valor *missing* es reemplazado por la media de todos los valores disponibles. Este modelo tiene el problema de no representar la distribución real del fenómeno ya que existirán muchos casos del valor de la media que no se ajustan a la realidad (Särndal y Lundström 2005). Para reducir esto se pueden realizar imputaciones por media para grupos concretos lo que reduciría este problema.
  Este método tiene la ventaja de ser muy sencillo y no necesitar información auxiliar pero solamente arrojaría resultados satisfactorios a la hora de calcular medias y totales poblacionales pero en ningún caso obtendríamos microdatos ajustados a la realidad. Por otro lado, la existencia de outliers afecta de manera muy negativa a esta técnica ya que las imputaciones se alejaran mas todavía de la distribución real.

- **Imputación por razón**. La imputación por razón tiene en cuenta una sola variable auxiliar y asume que esta variable es proporcional a la variable de estudio. El proceso de estimación consiste en calcular la razón de la variable auxiliar y la variable de estudio sobre el conjunto de datos sin valores *missing*. Una vez calculada esta razón se multiplica por el valor de la variable auxiliar de los valores *missing* y así se obtiene la imputación. Esta estimación será mejor cuanta mayor linealidad

exista entre la variable de estudio y la auxiliar. En este caso al igual que en la imputación por la media se pueden calcular razones para subgrupos dentro de la población si se dispone de más información. El inconveniente es que se necesita información sobre una variable auxiliar en aquellos que han respondido y en los que no lo han hecho. Para la obtención de esta información es habitual hacer uso de los registros estadísticos disponibles o otras operaciones estadísticas sobre la misma muestra.

- **Imputación por regresión**. Esta técnica es una generalización de las dos anteriores para un conjunto de variables auxiliares $x_1, ..., x_n$. El modelo mas simple es el de la regresión lineal que tiene la siguiente expresión:

$$y = \alpha + \beta_1 x_1 + ... + \beta_n x_n + \epsilon$$

donde α es el parámetro de la constante y $\beta_1, ..., \beta_n$ serán los parámetros desconocidos de cada una de las variables auxiliares, ϵ será el error e y es la variable de estudio. Normalmente la estimación de estos parámetros se realiza mediante mínimos cuadrados ordinarios.
El valor estimado se puede obtener tanto sumando el error como no haciéndolo. Sumar el error no es necesario cuando el objetivo de la imputación es obtener medias o totales poblacionales pero si se quiere observar la variabilidad es conveniente añadir el error (véase Waal, Pannekoek y Scholtus 2011, sección 7.3). El modelo sin error siempre tendrá el mismo resultado y por lo tanto es determinista pero la regresión añadiendo el componente de error será determinista si la forma de escoger el error lo es, de lo contrario las imputaciones serán estocásticas.
De esta fórmula se derivan las dos imputaciones explicadas anteriormente, por la media y por razón.
Estas técnicas serán más precisas cuanto mayor sea la relación entre las variables y normalmente se especifica una jerarquía en las técnicas para usar la más adecuada en cada unidad. Por ejemplo, si se ha comprobado que la técnica más precisa es la imputación por regresión y la siguiente más precisa la imputación por la media, se imputaría mediante regresión todas aquellas unidades que tengan información auxiliar disponible y para las que no lo tuvieran se usaría la imputación por la media (Särndal y Lundström 2005).

Imputación por donante *Hot deck*.
Esta técnica de imputación consiste en seleccionar a un 'donante' para la asignación del valor al receptor (Andridge y R. A. Little 2010). La selección del donante se puede hacer de diversos métodos pero el objetivo es obtener un donante lo mas similar posible para que la imputación sea mas precisa. De esta forma, una vez seleccionado el donante el proceso consistirá en imputar el valor del ítem del donante en el recipiente. Estas técnicas tienen la desventaja de que todos los donantes son parte de las observaciones y esto conlleva a asumir que no existen diferencias entre las personas que responden y las que no (Särndal y Lundström 2005).

Andridge y R. A. Little 2010 definen el termino *hot deck* como el uso de un donante

disponible en el mismo conjunto de datos que los valores ausentes y se contrapone a la imputación *cold deck* que consiste en el uso de conjuntos de datos diferentes para la selección del donante como por ejemplo periodos anteriores. Entre los métodos de imputación *hot deck* más usados destacan:

- **Imputación *hot deck* aleatoria y secuencial**.
  Es el método más simple de imputación usando un donante. El proceso consiste en seleccionar un donante al azar de las observaciones y usarlo para imputar un valor al recipiente. Este método tiene la ventaja de no requerir información auxiliar. En el caso de tener información adicional, como por ejemplo, la pertenencia a subgrupos de población se podría restringir la selección aleatoria del donante a ese grupo concreto. Si en vez de realizar una asignación aleatoria del donante se asigna la unidad más próxima en el registro con las características deseadas será imputación *hot deck* secuencial.
  En ambos métodos la desviación típica de los totales y la media aumentará ya que siempre existe la posibilidad de que un outlier sea el donante. Las estimaciones hot deck serán insesgadas únicamente para cuando los valores *missing* son MCAR que es poco probable que ocurra por lo que se recomienda reducirlo utilizando la información auxiliar (R. Little y Rubin 2002).

- **Imputación por el vecino más cercano**.
  Esta imputación se diferencia de las anteriores en que en vez de seleccionar una unidad con las mismas características se selecciona una unidad que minimice una función de distancia previamente definida. La función general de distancia usada es la distancia de Minkowski (Waal, Pannekoek y Scholtus 2011).
  En la versión mas simple donde solamente tenemos una variable auxiliar el donante será aquel para el que la diferencia en esa variable sea mínima. En el caso de tener varias variables auxiliares el donante seria aquel que menor suma de todas las distancias tuviera pero nótese que primero las variables se deben estandarizar o usar distancias relativas como la de Mahalanobis. A cada una de estas variables se le puede aplicar una ponderación para así dar mas o menos importancia a las variables que se deseen.
  La desventaja de esta técnica es que la distancia no puede calcularse de la misma forma para variables categóricas y numéricas pero esto puede solucionarse usando las variables categóricas para crear subgrupos homogéneos donde después extraer el donante (Waal, Pannekoek y Scholtus 2011).

Imputación por parte del experto.
Este método de imputación es el menos estadístico ya que consiste en que el o la responsable del producto estadístico determine el valor missing usando su capacidad analítica y toda la información de la que disponga.

24.4 Estrategia de depuración e imputación

La depuración de datos a menudo se realiza como una secuencia de distintos pasos de procesos de detección y/o corrección. Para finalizar este tema veamos una descripción global de una estrategia de depuración. Esta estrategia se representa en la Figura 24.1, que consiste en los siguientes cinco pasos.

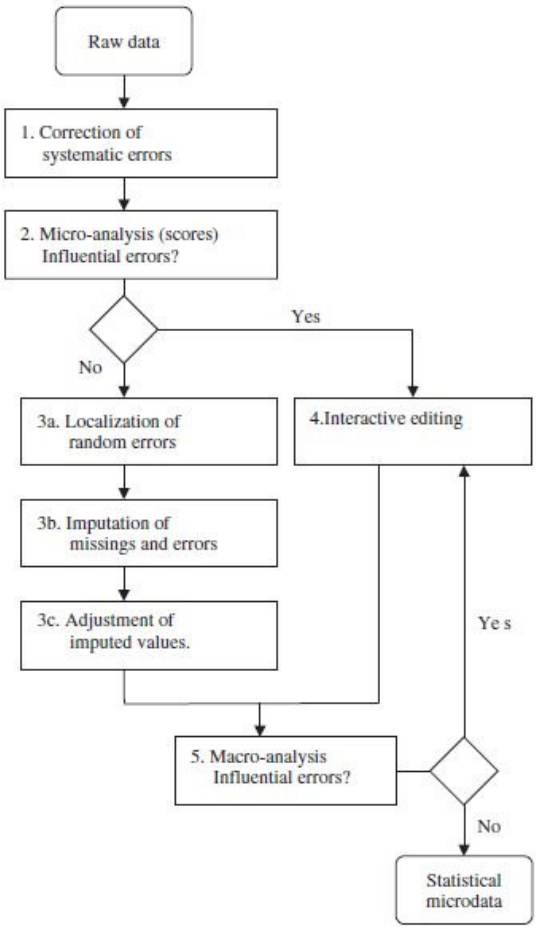

Figura 24.1: Estrategia general de depuración e imputación

1. *Tratamiento y corrección de errores sistemáticos*. Consiste en identificar y corregir los errores sistemáticos que son evidentes y fáciles de tratar con suficiente fiabilidad. Se puede hacer automáticamente con, virtualmente, ningún coste, y por tanto mejorar tanto la eficiencia como la calidad del proceso de depuración.

2. *Microselección*. Selecciona para su tratamiento interactivo registros que contienen errores influyentes que no pueden ser tratados de manera automática con suficiente fiabilidad. Por tanto serán controlados tanto manualmente (por expertos) como automáticamente (con edits especializados y algoritmos de depuración).

En este paso los datos se dividen en dos flujos: uno crítico y otro no crítico, usando técnicas de depuración selectiva. Para saber en qué medida un registro puede contener errores influyentes se puede usar una función *score*. Esta función se construye de forma que los registros con *scores* más altos se consideran como los que contienen efectos importantes sobre las estimaciones de los parámetros objetivo. Para ello se establece un umbral y todos los registros con *scores* por encima del umbral se revisan manualmente, mientras que los que estén por debajo se tratan de forma automática.

3. *Depuración automática*. Emplea los procedimientos automáticos de detección y corrección automática de errores a los registros que no son seleccionados para la depuración interactiva del paso 2. El primer paso en el tratamiento automático de errores es la localización de errores. Como los errores sistemáticos ya se han eliminado, los errores que todavía existen en este momento son aleatorios. Una vez que los errores duros se han definido y programado es fácil comprobar si los valores de un registro son inconsistentes en el sentido de que algunos de estos edits no se verifican. Sin embargo, no es tan obvio el decidir qué valores son erróneos en un registro inconsistente.

   A continuación se imputan los datos que faltan de manera automática. El mejor método de imputación para una situación particular dependerá de las características del conjunto de datos y su finalidad. En muchos casos los edits no son tenidos en cuenta por el método de imputación. Como consecuencia, los valores imputados pueden ser inconsistentes con las validaciones. Este problema se puede resolver introduciendo una fase de corrección en la cual se hacen ajustes en los valores imputados de forma que los registros verifiquen los edits y los ajustes sean lo más pequeños posible.

4. *Depuración interactiva*. Se aplica la depuración interactiva a la minoría de registros con errores influyentes. Los errores importantes en empresas grandes que tienen una gran influencia sobre los agregados que se publican y para los cuales no existen modelos de imputación acurados no se consideran adecuados para los procedimientos genéricos de depuración automática. Estos registros son tratados por expertos en un proceso llamado depuración interactiva.

5. *Macrodepuración*. Selecciona registros con errores influyentes usando métodos basados en técnicas de detección de *outliers* y otros procesos que hacen uso de toda o de una gran fracción de las respuestas. Los pasos anteriores usan todos métodos de microdepuración. Estos procesos de microdepuración se pueden realizar desde el principio de la fase de recogida de datos, tan pronto como los registros están disponibles.

   Por contra, las técnicas de macrodepuración usan información de otros registros y sólo se pueden usar si una gran parte de los datos ya se ha recogido o imputado.

> Las técnicas de macrodepuración también son técnicas de depuración selectiva en el sentido de que aspiran a prestar atención únicamente a posibles valores erróneos influyentes.

Aunque los procesos automáticos se usan con frecuencia para errores de poca importancia, elegir los métodos más adecuados de detección de errores e imputación es muy importante. Si se usan métodos inapropiados, especialmente para grandes cantidades de errores aleatorios y/o falta de respuesta, se puede introducir sesgo adicional.

Más aún, a medida que mejora la calidad de los métodos de localización automática de errores y de imputación, se pueden asignar más registros al tratamiento automático en el paso 3 y menos registros son seleccionados para el paso de depuración interactiva, que resulta mucho más costoso y consume mucho más tiempo.

El flujo de procesos sugerido en la Figura 24.1 es simplemente una posibilidad. Dependiendo del tipo de encuesta, de los recursos disponibles y de la información auxiliar, el flujo de procesos puede ser diferentes. No todos los pasos se realizan siempre y algunos de los pasos puede ser diferente.

Para encuestas sociales, por ejemplo, la depuración selectiva no es muy importante porque las contribuciones de los individuos al total publicado no son muy diferentes, al contrario de lo que ocurre con la contribución de las empresas pequeñas y grandes en una encuesta económica. A menudo, en las encuestas sociales, debido a la falta de edits duros, el principal tipo de error detectable es la falta de respuesta.

La UNECE ha desarrollado el Generic Statistical Data Editing Model (GSDEM) (UNECE 2019) como una referencia para todos los estadísticos oficiales entre cuyas actividades se incluya la depuración de datos. El GSDEM incluye las estrategias de depuración e imputación bajo distintos escenarios, encuestas sociales, encuestas económicas (coyunturales y estructurales), censos u operaciones basadas en la integración de datos.

Bibliografía

A. Wallgren, B. Wallgren (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley.

Andridge, R.R. y R.J. A. Little (2010). "Review of Hot Deck Imputation for Survey Non-response". En: *International Statistical Review* 78, págs. 40-64.

Couper, M.P., R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. NIchols y J.M. O'Reilly, eds. (1998). *Computer assisted survey information collection*. Wiley.

Eurostat (2014). *Handbook on Methodology of Modern Business Statistics*. URL: https://ec.europa.eu/eurostat/cros/content/imputation_en.

Fellegi, I.P. y D. Holt (1976). "A systematic approach to automatic edit and imputation". En: *J. Amer. Stat. Assoc.* 71, págs. 17-35.

Granquist, L. (1984). "Data Editing and its Impact on the Further Processing of Statistical Data". En: *Workshop on Statistical Computing, Budapest*.
– (1995). "Improving the Traditional Editing Process". En: Wiley, págs. 385-401.
– (1997a). "Macro-editing: a review of some methods for rationalizing the editing of survey data". En: *Statistical data editing: methods and techniques*.
– (1997b). "On the current best methods document: edit efficiciently". En: *UN/ECE Work Session on Statistical Data Editing* WWP. 30, págs. 1-8.
– (1997c). "The new view on editing". En: *International Statistical Review* 65, págs. 381-387.
Granquist, L. y J.G. Kovar (1997). "Editing of survey data: how much is enough?" En: Wiley, págs. 415-435.
Leeuw, E.D. de (2005). "To mix or not to mix data collection modes in surveys". En: *Journal of Official Statistics* 21, págs. 233-255.
Little, R.J.A. y D.B. Rubin (2002). *Statistical analysis with missing data*. 2nd. Hoboken: Wiley.
Nordbotten, S. (1955). "Measuring the Error of Editing the Questionnaires in a Census". En: *Journal of the American Statistical Association* 50, págs. 364-369.
Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.
Särndal, C.-E. y S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.
Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
UNECE (2019). *Generic Statistical Data Editing Model GSDEM*. Página visitada el día 15 de septiembre de 2021. URL: https://statswiki.unece.org/display/sde/GSDEM.
Waal, T. de, J. Pannekoek y S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Amsterdam: Wiley.
Whitridge, P. y J. Kovar (1990). "Applications of the generalised edit and imputation system at Statistics Canada". En: *Proceedings of the Section on Survey Research Methods*, págs. 105-110.

Tema 25

Metadatos de la producción estadística. I. GSBPM. Introducción. El modelo. Relaciones con otros modelos y estándares. Niveles 1 y 2 del GSBPM. Descripciones de fases y subprocesos (fases 1 a 3).

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

UNECE (2019c). *The Generic Statistical Business Process Model v5.1*. URL: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

25.1 Introducción

El *Generic Statistical Business Process Model* (GSBPM) (UNECE 2019c) es un modelo de procesos de negocio[1] que describe y define de manera genérica el conjunto de tareas de producción necesarias para elaborar estadísticas oficiales. Proporciona un marco estándar y terminología armonizada para ayudar a las organizaciones estadísticas a modernizar sus procesos de producción estadísticos, así como a compartir métodos y componentes del proceso.

El GSBPM también se puede usar para integrar estándares de datos y de metadatos, como una plantilla para documentar el proceso, para armonizar infraestructuras informáticas estadísticas y para proporcionar un marco para la evaluación y mejora de la calidad del proceso.

Este tipo de modelo estandarizado de producción surgió a principios del sigo XXI como una necesidad de la comunidad internacional en Estadística Oficial para modernizar e

[1] Traducimos *business process* como proceso de negocio; también se encuentra a veces como proceso de trabajo, proceso operativo o proceso de empresa, con ligeras variantes.

industrializar la producción de estadísticas oficiales por parte de los institutos nacionales de estadística. La versión actual del GSBPM es la 5.1 y está coordinada con las versiones 1.2 del *Generic Statistical Information Model* (GSIM) (UNECE 2019b) y la versión 1.2 del *Generic Activity Model for Statistical Organisations* (GAMSO) (UNECE 2019a).

25.2 El modelo

Un proceso de negocio estadístico es una sucesión de actividades y tareas que convierten datos de entrada en información estadística que responde a las necesidades de los usuarios. Esta información estadística está constituida por datos y metadatos presentados de diversos modos (tablas, mapas, gráficos, notas de prensa, etc.).

El GSBPM debe ser usado e interpretado con flexibilidad. ***No es un marco rígido en el cual todos los pasos de producción deben seguir un orden estricto***, en cambio identifica los posibles pasos del proceso estadístico y las inter-dependencias entre ellos.

Aunque la presentación del GSBPM sigue la secuencia lógica de pasos en la mayoría de los procesos estadísticos, los elementos del modelo (funciones de negocio) pueden ejecutarse en diferente orden dependiendo de la operación estadística en concreto. También, algunas tareas se repetirán un número de veces formando bucles iterativos, particularmente en las fases Procesar y Analizar tal y como se puede observar en la Figura 25.1.

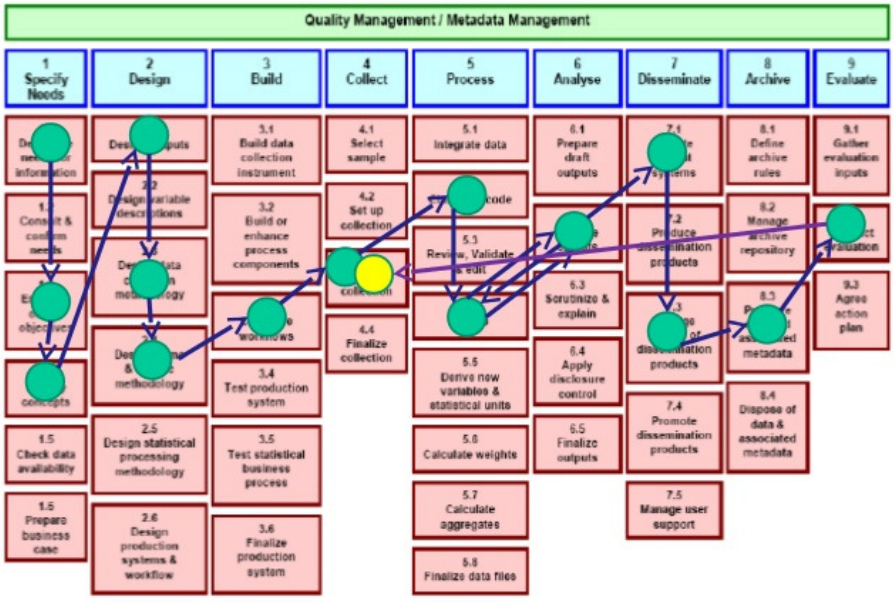

Figura 25.1: El GSBPM no es un estándar lineal.

El GSBPM se debe, por tanto, concebir más como una matriz, a través de la cual hay

muchas posibles rutas. En este sentido, el GSBPM aspira a ser lo suficientemente genérico como para ser ampliamente aplicable a muchos tipos de operaciones estadísticas, promoviendo así un punto de vista estandarizado del proceso estadístico en los institutos de estadística, sin llegar a ser ni demasiado restrictivo ni demasiado abstracto y teórico.

25.2.1 La estructura

El GSBPM abarca tres niveles:

- Nivel 0, el proceso estadístico;
- Nivel 1, las ocho fases del proceso estadístico;
- Nivel 2, los sub-procesos de cada fase.

Los niveles 1 y 2 están representados en la Figura 25.5.

El GSBPM también reconoce varios procesos globales[2] que se emplean a lo largo de las ocho fases. Éstos se pueden agrupar en dos categorías, los que tienen un componente estadístico y los procesos de soporte que son más generales y se pueden emplear en cualquier tipo de organización. Los del primer grupo se consideran más importantes en el contexto de este modelo, sin embargo también se deberán reconocer que los del segundo grupo (incluidos con detalle en el GAMSO (UNECE 2019a)) tienen (a menudo de manera indirecta) impacto en varias partes del modelo.

Los procesos globales con una componente estadística incluyen los siguientes. Los cuatro primeros están más relacionados con el modelo.

- Gestión de calidad.- Este proceso incluye la evaluación de la calidad y mecanismos de control del proceso. Reconoce la importancia de la evaluación y el *feedback* a lo largo de todo el proceso estadístico;
- Gestión de metadatos.– Los metadatos son generados/reusados y procesados en cada fase, de donde se deriva una fuerte necesidad de disponer de un sistema de gestión de metadatos que asegure que los metadatos apropiados permanecen enlazados con los datos a lo largo de todo el GSBPM. Esto incluye consideraciones independientes del proceso tales como las figuras de custodio y responsable de los metadatos, su calidad, las reglas de archivo, preservación[3], conservación[4] y eliminación;
- Gestión de datos - Esto incluye consideraciones independientes del proceso tales como la seguridad general de los datos, las figuras de custodio y responsable

[2] Traducimos *overarching* como global.

[3] Por preservación (*preservation*) se entiende el acto de conservar y mantener tanto la seguridad como la integridad de los datos. Se lleva a cabo mediante actividades formales gobernadas por políticas, regulaciones y estrategias dirigidas a proteger y prolongar la existencia y autenticidad de los datos y metadatos (Wikipedia 2021a).

[4] Por conservación (*retention*) definimos las políticas de gestión de datos y registros permanentes para cumplir los requisitos legales y de negocio de archivación de datos (Wikipedia 2021b).

de los datos, la calidad de los datos, las reglas de almacenamiento, preservación, conservación y eliminación;

- Gestión de los datos de procesos.– Esto incluye las actividades de registro, sistematización y uso de los datos de la implementación del proceso de negocio.
- Gestión del conocimiento.– Asegura que los procesos estadísticos son repetibles, principalmente gracias al mantenimiento de la documentación del proceso;
- Gestión del marco estadístico.– Incluye desarrollar estándares, por ejemplo metodologías, conceptos y clasificaciones se utilizan de forma generalizada en múltiples procesos;
- Gestión del programa estadístico.– Incluye la monitorización sistemática y revisión de necesidades incipientes de información así como fuentes de datos nuevas y que cambian entre todos los dominios estadísticos. Puede dar como resultado la definición de nuevos procesos estadísticos o el rediseño de los existentes;
- Gestión de proveedores.– Incluye la gestión de la carga de procesos trasversales, y temas como la caracterización y la gestión de la información de contacto (y por tanto está particularmente enlazada con procesos estadísticos que mantienen registros);
- Gestión de los usuarios.– Incluye actividades generales de marketing, promoción de los conocimientos estadísticos, y encargarse del *feedback* de usuarios no específicos.

Procesos globales más generales incluyen:

- Gestión de los recursos humanos;
- Gestión económica;
- Gestión de proyectos;
- Gestión del marco legal;
- Gestión del marco organizativo;
- Planificación estratégica.

25.2.2 Aplicabilidad

El objeto del GSBPM es su aplicación a todas las actividades llevadas a cabo por los productores de estadísticas oficiales, tanto a nivel nacional como a nivel internacional, con las que se obtienen resultados estadísticos. Se ha diseñado para ser independiente de la fuente de datos, por lo que puede ser utilizado para la descripción y evaluación de la calidad de procesos basados en encuestas, censos, registros administrativos, y otras fuentes no estadísticas o combinadas[5].

[5]Esta afirmación, que aparece en el GSBPMv5.1, es discutible. Existen proyectos internacionales en el Sistema Estadístico Europeo que proponen arquitecturas de producción estadística que generalizan la colección de funciones de negocio del GSBPM para abarcar necesidades de la producción derivadas del uso de *Big Data* (véase p.ej. ESSnet on Big Data 2021).

Mientras que los procesos estadísticos típicos incluyen la recogida y el procesamiento de los datos para producir resultados estadísticos, el GSBPM también se puede utilizar en casos en los que los datos existentes se revisan o que series de datos son recalculadas, tanto como resultado de una mejora en las fuentes de datos, como por un cambio en la metodología estadística. En estos casos, los datos de entrada son las estadísticas previamente publicadas, que son entonces procesadas y analizadas para producir resultados revisados. En tales casos, es probable que varios subprocesos y posiblemente algunas fases (particularmente las iniciales) sean omitidas. De forma similar, el GSBPM también se puede utilizar en procesos tales como la síntesis de Cuentas Nacionales y los típicos procesos de organizaciones estadísticas internacionales.

Además de usarse para procesos de los que se obtienen estadísticas, el GSBPM también se puede utilizar en el desarrollo y mantenimiento de registros estadísticos, donde los inputs son similares a los que se usan para la producción estadística (aunque típicamente poniendo el foco en datos administrativos), y los resultados son típicamente marcos u otras extracciones de datos, que serán entonces usados como inputs para otros procesos.

El GSBPM debe verse como un instrumento lo suficientemente flexible como para ser utilizado en todos los supuestos anteriores.

25.2.3 El uso del GSBPM

El GSBPM es un modelo de referencia. Está previsto que el GSBPM pueda ser utilizado por organizaciones a distintos niveles. Una organización puede elegir implementar el GSBPM de forma directa o usarlo como la base para desarrollar una adaptación específica del modelo.

Por ejemplo, el INE ha desarrollado un tercer nivel del GSBPM adaptado a las necesidades del Sistema Estadístico Nacional en España (INE 2015). Se puede usar en algunos casos sólo como un modelo al que las organizaciones se refieren en su comunicación interna o con otras organizaciones para clarificar discusiones. Los distintos escenarios para el uso del GSBPM son todos válidos.

Cuando las organizaciones hayan desarrollado adaptaciones específicas del GSBPM, pueden hacer algunas especializaciones al modelo para ajustarse a su contexto (véase p.ej. INE 2015). La evidencia hasta ahora sugiere que estas especializaciones no son suficientemente genéricas como para ser incluidas en el GSBPM.

En algunos casos puede ser apropiado agrupar algunos de los elementos del modelo. Por ejemplo, las fases uno a tres se puede considerar que corresponden a una única fase. En otros casos, de forma particular para implementaciones prácticas, puede surgir la necesidad de añadir uno o más niveles detallados a la estructura indicada a continuación para identificar de manera separada diferentes componentes de los subprocesos.

25.3 Relaciones con otros modelos y estándares

Desde que se publicó el GSBPM, se han desarrollado varios modelos bajo el auspicio del HLG-MOS[6] para ayudar en la modernización de las estadísticas oficiales. De forma colectiva, se llama los modelos "ModernStats". Los siguientes párrafos muestran los modelos ModernStats que están más relacionados con el GSBPM. Esta relación se muestra en la Figura 25.2

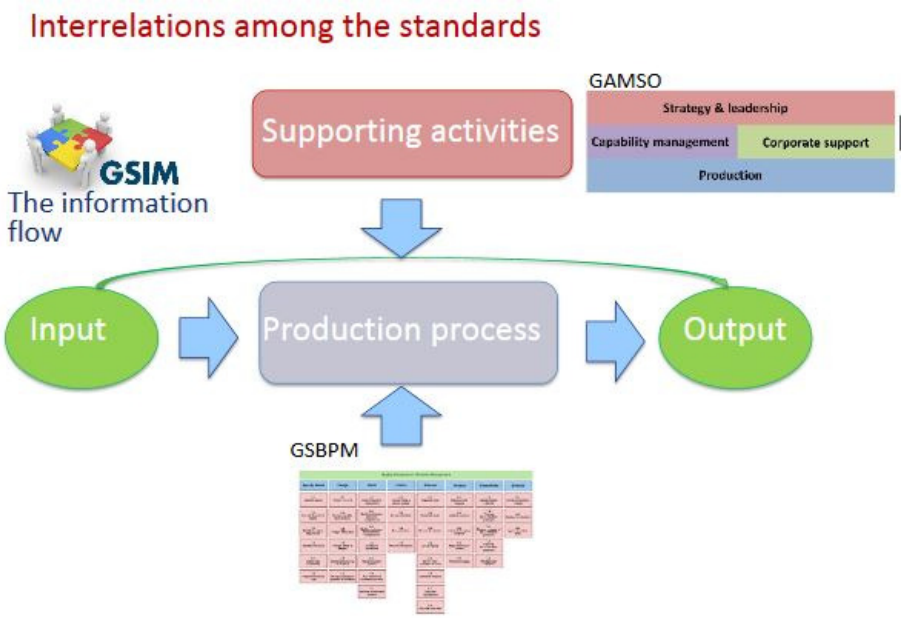

Figura 25.2: Relación entre GSIM, GSBPM y GAMSO.

25.3.1 GAMSO

El GAMSO (UNECE 2019a) describe y define actividades que tienen lugar en una organización estadística. Extiende y complementa al GSBPM añadiendo actividades necesarias para la producción estadística (es decir actividades en las áreas de estrategia y dirección, desarrollo de capacidades y apoyo corporativo). En el GSBPM v5.0, algunas de estas actividades estaban incluidas como procesos generales. Las actividades que no están directamente relacionadas con la producción de estadísticas y/o son gestionadas a nivel corporativo o estratégico están ahora incluidas en el GAMSO (p.ej. gestión de los recursos humanos o las actividades de gestión de la calidad que se llevan a cabo a nivel corporativo como el desarrollo de un marco de calidad).

El GAMSO describe actividades – es decir, lo que las organizaciones estadísticas hacen. Incluye descripciones a alto nivel de estas actividades. Por un lado, el GSBPM se centra en el proceso de producción– describe más detalladamente cómo las organizaciones estadísticas llevan a cabo la actividad de producción estadística.

[6] *High-Level Group for the Modernisation of Official Statistics.*

Como el GSBPM, el GAMSO busca proporcionar un vocabulario común y un marco que fomente las actividades internacionales de colaboración. Se obtendrá un mayor valor del GAMSO si se aplica de forma conjunta con el GSBPM.

25.3.2 GSIM

El GSIM (UNECE 2019b) es un marco de referencia para la **información estadística**, diseñado para jugar un papel importante en la modernización y racionalización de estadísticas oficiales tanto a nivel nacional como internacional. Facilita descripciones genéricas de la definición, gestión y uso de los datos y metadatos a lo largo del proceso de producción estadística. Proporciona un conjunto de objetos de información estandarizados y consistentemente descritos, que son los inputs y outputs en el diseño y producción de estadísticas.

El GSIM ayuda a explicar las relaciones significativas entre las entidades involucradas en la producción estadística y se puede usar para orientar el desarrollo y uso de estándares o especificaciones de implementación que resulten consistentes.

Como el GSBPM, el GSIM es uno de los pilares para la modernización de las estadísticas oficiales y alejarse del modelo actual de compartimentos estancos (*stovepipe*). El GSIM se ha diseñado para permitir enfoques innovadores en la producción estadística en la mayor medida posible; por ejemplo, en el área de la difusión, donde las demandas de agilidad e innovación están aumentando. También proporciona apoyo a aproximaciones vigente de la producción estadística.

El GSIM y el GSBPM son modelos complementarios para la producción y gestión de la información estadística. Como se muestra en la Figura 25.3, el GSIM ayuda a describir los subprocesos del GSBPM definiendo los objetos de información que fluyen entre ellos, que se crean en ellos, y que son usados por ellos para producir estadísticas oficiales. Los inputs y outputs se pueden definir en término de los objetos de información, y se formalizan en el GSIM.

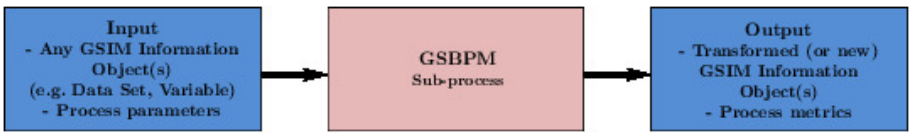

Figura 25.3: Relación entre GSIM y GSBPM en un paso de producción estándar.

Por tanto, se obtendrá un mayor valor del GSIM si se aplica de forma conjunta con el GSBPM. Del mismo modo, se obtendrá un mayor valor del GSBPM si se aplica de forma conjunta con el GSIM. Sin embargo, es posible (aunque no ideal) aplicar el uno sin el otro.

De forma similar, ambos modelos apoyan la implementación de la *Common Statistical Production Architecture* (CSPA) (UNECE 2021), pero se puede aplicar con independencia

de si se usa o no este marco de arquitectura de producción.

Aplicar de manera conjunta el GSIM y el GSBPM puede facilitar la construcción de sistemas eficientes gestionados con metadatos, y ayudar a armonizar infraestructuras informáticas estadísticas.

Una versión más desglosada de la Figura 25.3 se representa en la Figura 25.4, donde se observan tanto el flujo de datos como el flujo de metadatos[7].

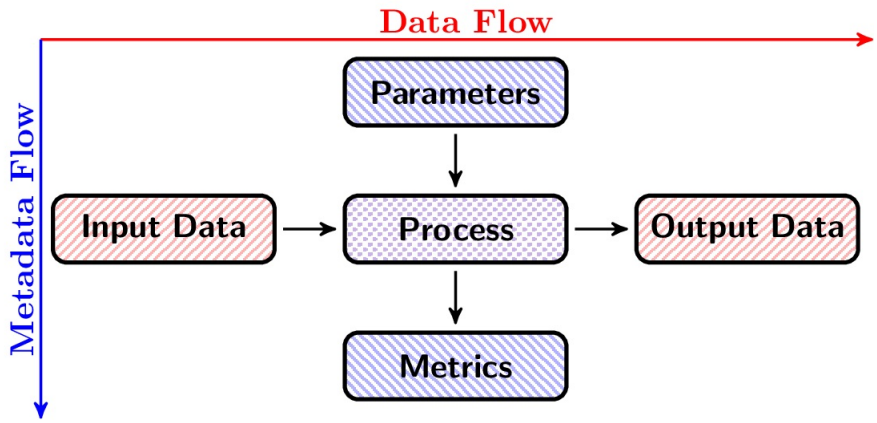

Figura 25.4: Relación entre GSIM y GSBPM en un paso de producción estándar desglosado.

25.4 Niveles 1 y 2 del GSBPM

Los niveles 1 y 2 del GSBPM están representados matricialmente en la Figura 25.5.

[7]Esta Figura está tomada de una conferencia de M. van der Loo en el INE en el año 2018 (véase Loo 2021, para sus referencias).

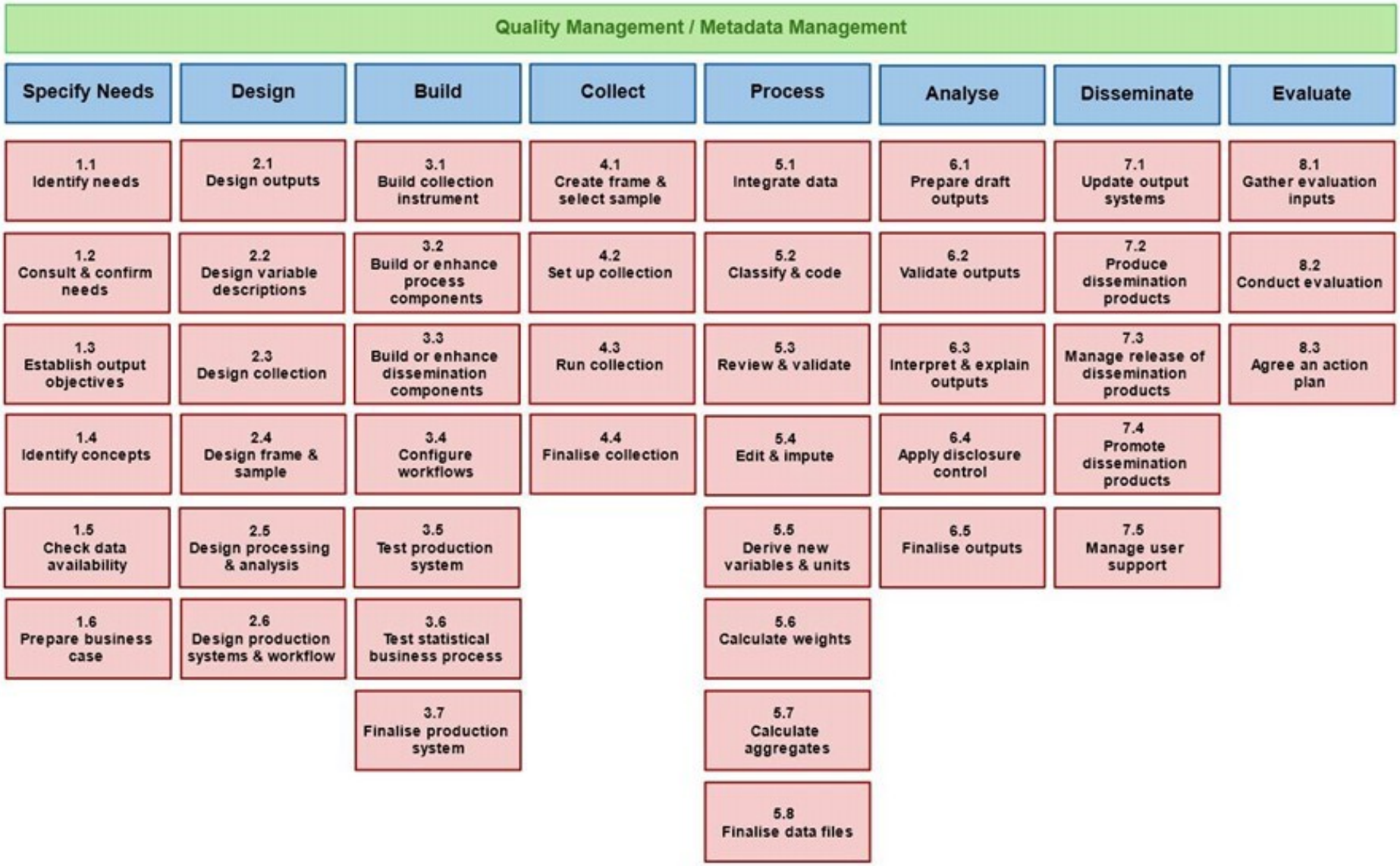

Figura 25.5: Niveles 1 y 2 del GSBPM

25.5 Descripciones de fases y subprocesos (fases 1 a 3)

A continuación se define cada fase, identificando los subprocesos dentro de cada fase, y describiendo sus contenidos.

1. Especificar necesidades

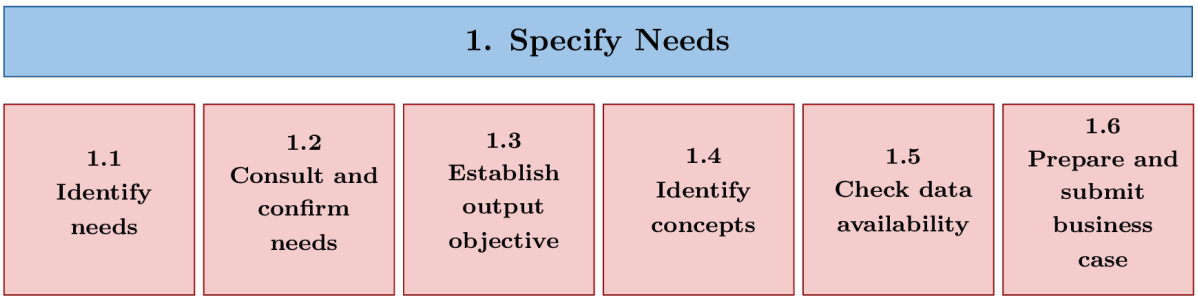

Figura 25.6: Fase 1 del GSBPM

Esta fase se desencadena cuando se detecta la necesidad de una nueva estadística o se inicia la revisión de una estadística existente como consecuencia de algún *feedback*. Incluye todas las actividades asociadas con la puesta en contacto con los usuarios para identificar de forma detallada sus necesidades estadísticas, proponiendo soluciones y preparando un proyecto que resuelva estas necesidades.

En esta fase la organización:

- identifica las necesidades de estadísticas;
- confirma, en mayor detalle, las necesidades de las distintas partes;
- establece los objetivos de los resultados estadísticos;
- identifica los conceptos y variables relevantes para los que se piden los datos;
- comprueba hasta qué punto las fuentes de datos actuales permiten alcanzar estas necesidades;
- prepara la documentación para la elaboración y justificación de la necesidad y viabilidad de un nuevo proyecto para conseguir la aprobación para producir las estadísticas.

Esta fase se divide en seis subprocesos. Estos son generalmente secuenciales, de izquierda a derecha, pero también pueden tener lugar en paralelo y pueden ser iterativos. Los subprocesos son:

1.1. Identificar necesidades

Este subproceso incluye la investigación inicial e identificación de qué estadísticas son necesarias y qué se necesita de las estadísticas. Puede estar provocada por una nueva

solicitud de información o un cambio, como una reducción del presupuesto. Planes de acción provenientes de evaluaciones en iteraciones previas del proceso o de otros procesos relacionados pueden proporcionar un input a este subproceso. También incluye consideraciones sobre prácticas entre otras organizaciones estadísticas (nacionales e internacionales) que producen datos similares y, en particular, los métodos usados por esas organizaciones. Puede incluir consideraciones sobre las necesidades específicas de distintos tipos de usuarios, como los discapacitados, o distintos grupos étnicos.

1.2. Consultar y confirmar necesidades

Este subproceso se centra en la consulta con las distintas partes interesadas y usuarios y confirma detalladamente las necesidades para producir las estadísticas en consideración. Se requiere un buen entendimiento de las necesidades de los usuarios para que la organización estadística sepa no solo qué es necesario publicar, sino también cuándo, cómo, y quizá, lo más importante, por qué. Para la segunda y posteriores iteraciones de esta fase, el principal objetivo será determinar si las necesidades previamente identificadas han cambiado. Este buen entendimiento de las necesidades de los usuarios es la parte crítica de este subproceso.

1.3. Establecer objetivos de los resultados

Este subproceso identifica los resultados estadísticos que son necesarios para alcanzar las necesidades de los usuarios identificadas en el subproceso 1.2 (Consultar y confirmar necesidades). Incluye ponerse de acuerdo en la idoneidad de los resultados propuestos y sus medidas de calidad con los usuarios. Los marcos legales (p.ej. relacionados con la confidencialidad) y los recursos disponibles pueden ser limitaciones a la hora de establecer los objetivos.

1.4. Identificar conceptos

Este subproceso clarifica los conceptos requeridos que serán medidos durante el proceso desde el punto de vista de los usuarios. En este punto los conceptos identificados pueden aún no estar alineados con estándares estadísticos existentes. Este alineamiento y la elección o definición de los conceptos y variables estadísticas que serán usadas tiene lugar en el subproceso 2.2.

1.5. Comprobar disponibilidad de datos

Este subproceso comprueba si las fuentes actuales de datos podrían alcanzar los requisitos de los usuarios y las condiciones bajo las cuales estarían disponibles, incluyendo cualquier restricción en su uso. Una evaluación de posibles alternativas normalmente incluiría investigar potenciales fuentes de datos como los registros administrativos u fuentes otras no estadísticas para determinar si se podrían usar con fines estadísticos. Una vez que se han analizado las fuentes, se prepara una estrategia para cubrir las lagunas restantes. Este subproceso también incluye una evaluación más general sobre el marco legal en el cual se recogerán y usarán los datos y se pueda así identificar pro-

puestas de cambios en la legislación existente o la introducción de un nuevo marco legal.

1.6. Elaborar documentación para la elaboración y justificación de la necesidad y viabilidad de un nuevo proyecto

Este subproceso elabora documentación con los resultados de los otros subprocesos de esta fase en forma de caso de uso (*business case*) para evaluar y aprobar la implementación del proceso estadístico nuevo o modificado. Esta documentación necesitaría ajustarse a los requisitos de quien tiene que dar el visto bueno, pero generalmente incluye elementos como:

- Una descripción del proceso 'Como-Está'[8] proceso (si ya existe), con información sobre cómo se producen las estadísticas actuales, señalando las ineficiencias y aspectos a tener en cuenta;
- La solución propuesta para el proceso 'Futuro'[9], detallando cómo el proceso estadístico se desarrollará para producir las estadísticas nuevas o revisadas;
- Una evaluación de los costes y beneficios así como cualquier restricción externa.

2. Diseñar

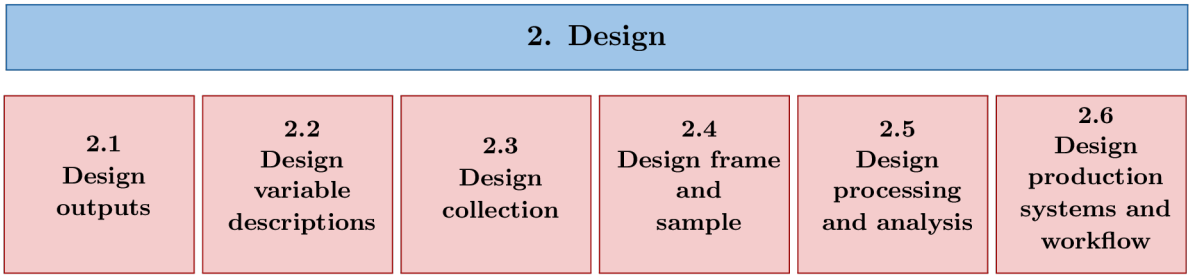

Figura 25.7: Fase 2 del GSBPM

Esta fase describe el desarrollo y actividades de diseño así como cualquier otro trabajo práctico de investigación asociado necesarios para definir los resultados estadísticos, conceptos, metodologías, instrumentos de recogida y procesos operacionales. Incluye todos los elementos del diseño necesarios para definir o refinar los productos estadísticos o servicios identificados en la documentación del caso de uso (subproceso 1.6.). Esta fase especifica todos los metadatos relevantes listos para su uso más tarde en el proceso estadístico, así como los procedimientos de garantía de calidad. Para resultados estadísticos producidos de forma periódica, esta fase normalmente tiene lugar la primera vez y cuando se identifiquen acciones de mejora en la fase de evaluación en alguna iteración previa.

Las actividades de diseño emplean estándares internacionales y nacionales con el fin de reducir la duración y el coste del diseño de proceso mejorando así la comparabilidad

[8] Traducimos *As-Is process* como proceso 'Como-Está'.
[9] Traducimos *To-Be process* como proceso 'Futuro'.

y usabilidad de los resultados. En este sentido, las organizaciones deben fomentar la reutilización o la adaptación de elementos de procesos existentes. Adicionalmente, los resultados de los procesos de diseño pueden formar la base de estándares futuros en la organización a nivel nacional o internacional.

Esta fase se divide en seis subprocesos, que son generalmente secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

2.1. Diseñar resultados

Este subproceso contiene el diseño detallado de los resultados estadísticos, productos y servicios que se producirán, incluyendo los trabajos de desarrollo y preparación de los sistemas e instrumentos usados en la fase 'Difundir' (proceso 8.). Los métodos de control del secreto estadístico, así como los procesos relacionados con el acceso a los resultados confidenciales, también se diseñan en este subproceso. Los resultados deberían ser diseñados de forma que sigan los estándares existentes siempre que sea posible, por lo que los inputs de este proceso pueden incluir metadatos de operaciones de recogida de datos similares o anteriores, estándares internacionales e información sobre prácticas en otras organizaciones estadísticas relativas al subproceso 1.1 (Identificar necesidades).

2.2. Diseñar descripciones de variable

Este subproceso define las variables estadísticas que se recogerán mediante los instrumentos de recogida, así como cualesquiera otras variables que se obtendrán a partir de ellas en el subproceso 5.5 (Derivar nuevas variables y unidades) y las clasificaciones estadísticas que se usarán. Se espera que se sigan, siempre que sea posible, los estándares nacionales e internacionales existentes. Puede ser necesario que este subproceso vaya en paralelo con el subproceso 2.3 (Diseñar recogida/obtención), mientras que la definición de las variables a recoger y la elección de los instrumentos de recogida pueden ser interdependientes hasta cierto punto. La preparación de la descripción de los metadatos de las variables recogidas y derivadas y las clasificaciones es una precondición necesaria para las fases posteriores.

2.3. Diseñar recogida

Este subproceso determina los métodos e instrumentos de recogida más adecuados. Las actividades de este subproceso variarán de acuerdo con el tipo de instrumentos de recogida necesarios, que pueden incluir entrevistas asistidas por ordenador, cuestionarios en papel, interfaces con datos administrativos y técnicas de integración de datos. Este subproceso incluye el diseño de los instrumentos de recogida, preguntas y plantillas de respuesta (junto con las variables y clasificaciones estadísticas diseñados en el subproceso 2.2 (Diseñar descripciones de variable)). También incluye el diseño de cualquier acuerdo formal relacionado con el suministro de datos, tal como los memorandos de acuerdo y la confirmación de la base legal para la recogida de datos. Este subproceso se ve favorecido por herramientas tales como librerías de preguntas

(para facilitar la reutilización de preguntas y atributos relacionados), herramientas de diseño de cuestionarios (para posibilitar una compilación rápida y fácil de preguntas en formatos adecuados para tests cognitivos) y modelos de acuerdos (para ayudar a estandarizar términos y condiciones). Este subproceso también incluye el diseño de sistemas de gestión del informante específicos del proceso.

2.4. Diseñar marco y muestra

Este subproceso solo se utiliza en procesos que involucren la recogida de datos basada en el muestreo. Este subproceso identifica y especifica la población de interés, define el marco muestral (y, donde sea necesario, el registro del que se deriva), y determina el criterio más apropiado de muestreo y la metodología (que puede incluir una enumeración completa). Las fuentes comunes para un marco muestral son los registros administrativos y estadísticos, los censos y la información de otras encuestas muestrales. Este subproceso describe cómo estas fuentes se pueden combinar si es necesario. Debe abordarse un análisis sobre si el marco cubre la población objetivo. Debe elaborarse un plan de muestreo. La muestra concreta se crea en el subproceso 4.1 (Crear marco y seleccionar muestra), usando la metodología especificada en este subproceso.

2.5. Diseñar procesamiento y análisis

Este subproceso diseña la metodología del proceso estadístico que se aplicará durante las fases 'Procesar' y 'Analizar'. Puede incluir especificaciones de rutinas para codificar, depurar, imputar, estimar, integrar, validar y finalizar conjuntos de datos.

2.6. Diseñar sistemas de producción y flujos de trabajo

Este subproceso determina los flujos de trabajo desde la recogida de datos hasta la difusión, con una visión total de todos los procesos necesarios en el conjunto del proceso de producción estadístico, y asegurando que todos juntos encajan de forma eficiente sin huecos o redundancias. A través del proceso se necesitan varios sistemas y bases de datos. Un principio general es la reutilización de procesos y tecnología entre muchos procesos estadísticos, por lo que las soluciones existentes en la producción (por ejemplo, servicios, sistemas y bases de datos) deberían de ser examinadas en primer lugar, para determinar si son adecuadas para este proceso específico, y, si se identifica alguna laguna, diseñar nuevas soluciones. Este subproceso también considera cómo el personal interactuará con los sistemas, y quién será responsable de qué y cuándo.

3. Desarrollar

Esta fase construye y prueba las soluciones de producción hasta que estén listas para su uso en el entorno real de producción. Los resultados de la fase 'Diseñar' apunta a la selección de procesos reutilizables, instrumentos, información y servicios que estén ensamblados y configurados en esta fase para crear el entorno operacional completo para llevar a cabo el proceso. Nuevos servicios se desarrollan como una excepción, creados como respuesta a los vacíos en el catálogo actual de servicios existentes tanto

dentro de la organización como externamente. Estos nuevos servicios se construyen para que puedan ser ampliamente reutilizados dentro de la arquitectura de producción estadística.

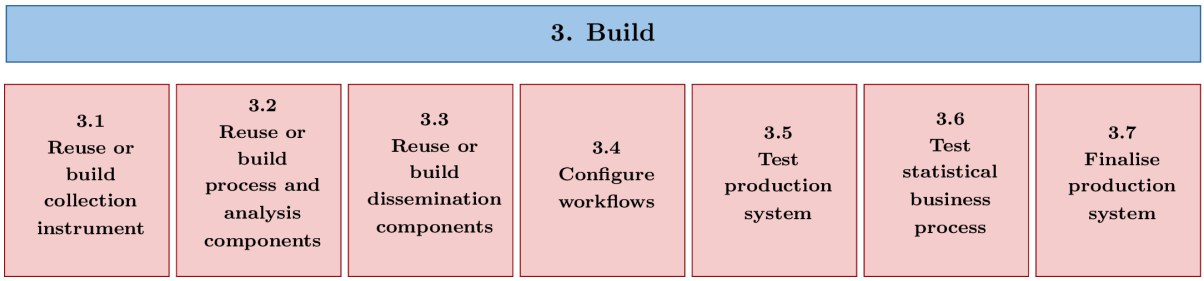

Figura 25.8: Fase 3 del GSBPM

Para los resultados estadísticos producidos de forma periódica, esta fase, más que en cada iteración, normalmente ocurre en la primera iteración y después de una revisión o un cambio metodológico o tecnológico.

Se descompone en siete subprocesos, que son generalmente secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo y pueden ser iterativos. Estos subprocesos son:

3.1. Desarrollar herramientas de recogida[10]

Este subproceso describe las actividades para construir los instrumentos de recogida que se usarán en la fase 'Recoger/Obtener'. El instrumento de recogida se genera o construye basado en las especificaciones de diseño creadas durante la fase 'Diseñar'. La recogida puede usar una o más formas de recogida, por ejemplo entrevistas personales o por teléfono; cuestionarios en papel, electrónicos o por web; centros de SDMX (*SDMX hubs*). Los instrumentos de recogida también pueden ser rutinas de extracción de datos usadas para combinar datos de conjuntos de datos estadísticos o administrativos ya existentes. Este subproceso también incluye preparar y probar los contenidos y funcionalidades del instrumento (por ejemplo, probar las preguntas en un cuestionario). Se recomienda considerar la conexión de los instrumentos de recogida con los sistemas de metadatos estadísticos, de forma que los metadatos se puedan recoger más fácilmente durante la fase de recogida. La conexión de los metadatos y los datos en el momento de recogida puede ahorrar trabajo en fases posteriores. La recogida de las métricas e indicadores de la recogida de datos (paradatos) también se considera importante en este subproceso.

3.2. Desarrollar o mejorar componentes de procesamiento

[10]Se toman los nombres de los subprocesos del GSBPM del estándar del INE (INE 2015).

Este subproceso describe las actividades para construir nuevos componentes y servicios o mejorar los existentes que son necesarios para las fases 'Procesar' y 'Analizar', tal y como están diseñados en la fase 'Diseñar'. Los servicios pueden incluir funciones y características para consolas de control (*dashboard*), servicios de información, funciones de transformación, entornos de flujos de trabajo y servicios de suministro y gestión de metadatos.

3.3. Desarrollar o mejorar componentes de difusión

Este subproceso describe las actividades para construir nuevos componentes y servicios o mejorar los existentes que son necesarios para la difusión de los productos estadísticos diseñados en el subproceso 2.1 (Diseñar resultados). Todos los tipos de componentes y servicios de difusión deben estar incluidos, desde los que se usan para producir publicaciones tradicionales en papel a los proporcionados por los servicios web, datos abiertos (*open data*) o accesos a microdatos.

3.4. Configurar flujos de trabajo

Este subproceso configura los flujos de trabajo, sistemas y transformaciones usadas en los procesos estadísticos, desde la recogida de datos a la difusión. Asegura que el flujo de trabajo especificado en el subproceso 2.6 (Diseñar sistemas de producción y flujo de trabajo) funciona en la práctica.

3.5. Probar sistemas de producción

Este subproceso está relacionado con la prueba de servicios montados y configurados así como con los flujos de trabajo relacionados. Incluye pruebas técnicas y la aprobación conjunta de los nuevos programas y rutinas, así como la confirmación de que las rutinas existentes en otros procesos estadísticos son adecuadas para su uso en este caso. Mientras que parte de esta actividad relacionada con la prueba de componentes y servicios individuales podría estar lógicamente enlazado con el subproceso 3.2 (Desarrollar o mejorar componentes de procesamiento), este subproceso también incluye la prueba de las interacciones entre servicios ensamblados y configurados y asegurar que las soluciones de producción funcionan como un conjunto coherente de procesos, información y servicios.

3.6. Probar procesos estadísticos

Este subproceso describe las actividades para gestionar pruebas de campo o pruebas piloto del proceso estadístico. Normalmente incluye una recogida a pequeña escala, para probar los instrumentos de recogida, seguido por el procesamiento y el análisis de los datos recogidos, para asegurar que el proceso estadístico funciona como se espera. Después del piloto, puede ser necesario volver a algún paso anterior y hacer ajustes en los instrumentos, sistemas o componentes. Para un proceso estadístico enorme, por ejemplo un censo de población, pueden ser necesarias varias iteraciones hasta que el proceso funciona de manera satisfactoria.

3.7. Finalizar sistema de producción

Este subproceso incluye las actividades para poner en marcha procesos y servicios montados y configurados, incluyendo servicios modificados y de nueva creación en producción listos para su uso. Las actividades incluyen:

- producir documentación sobre las componentes del proceso, incluyendo documentación técnica y manuales de usuarios;
- formación de los usuarios sobre cómo funciona el proceso;
- trasladar las componentes del proceso a un entorno de producción y asegurar que funcionan como se espera en tal entorno (esta actividad también puede ser parte del subproceso 3.5 (Probar sistema de producción)).

Bibliografía

ESSnet on Big Data (2021). *Work Package F on Process and Architecture*. Página visitada el día 15 de septiembre de 2021. URL: https://ec.europa.eu/eurostat/cros/content/WPF_Process_and_architecture_en.

INE (2015). *Estándar de documentación de procesos de producción de operaciones estadísticas del INE: Los informes estandarizados de los metadatos de proceso*. Página visitada el día 15 de septiembre de 2021. URL: https://www.ine.es/clasifi/estandar_procesos.pdf.

Loo, M. van der (2021). *Home Page*. URL: http://www.markvanderloo.eu/.

UNECE (2019a). *Generic Activity Model for Statistical Organizations*. URL: https://statswiki.unece.org/display/GAMSO/.

– (2019b). *Generic Statistical Information Model v1.2*. URL: https://statswiki.unece.org/display/gsim/.

– (2019c). *The Generic Statistical Business Process Model v5.1*. URL: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

– (2021). *Common Statistical Production Architecture*. URL: https://statswiki.unece.org/display/CSPA.

Wikipedia (2021a). *Data Preservation*. Página visitada el día 15 de septiembre de 2021. URL: https://en.wikipedia.org/wiki/Data_preservation.

– (2021b). *Data Retention*. Página visitada el día 15 de septiembre de 2021. URL: https://en.wikipedia.org/wiki/Data_retention.

Tema 26

Metadatos de la producción estadística. II. GSBPM. Descripciones de fases y subprocesos (fases 4 a 8). Procesos generales. Otros usos del GSBPM. Data Documentation Initiative (DDI), SDMX y comparación con el GSBPM.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía:

UNECE (2019c). *The Generic Statistical Business Process Model v5.1*. URL: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

26.1 Descripciones de fases y subprocesos (fases 4 a 8)

4. Recoger/Obtener[1]

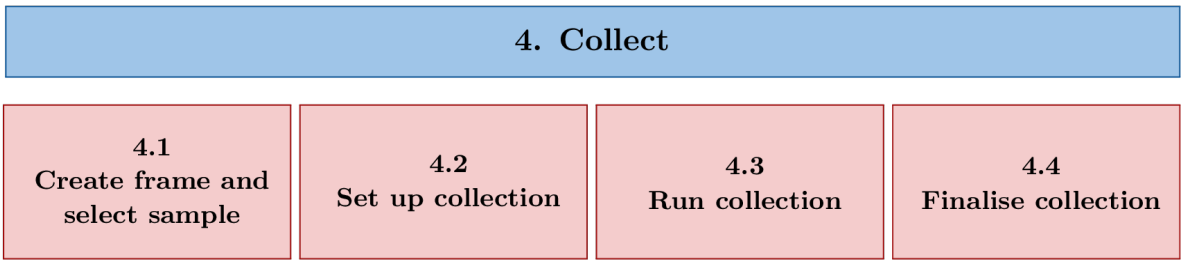

Figura 26.1: Fase 4 del GSBPM

[1]Esta fase se denomina Recoger/Obtener. En lo que sigue se utilizará únicamente la palabra recoger. Obtener se refiere a las operaciones estadísticas en las que los datos no se recogen directamente de las unidades de la muestra mediante cuestionarios, sino que se obtienen a partir de registros administrativos o de otras fuentes de datos.

Esta fase recoge o recolecta toda la información necesaria (datos y metadatos), usando diferentes métodos de recogida (incluyendo la extracción de registros estadísticos, administrativos y otros no estadísticos así como bases de datos) y los carga en un entorno adecuado para un procesamiento posterior. Mientras que esto puede incluir la validación de los formatos de los conjuntos de datos, no incluye ninguna transformación de los datos en sí mismos, ya que todo esto se hace en la fase 'Procesar' (fase 5.). Para resultados estadísticos producidos de forma periódica, esta fase ocurre en cada iteración.

La fase 'Recoger' se divide en cuatro subprocesos, que son generalmente secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo y pueden ser iterativos. Estos subprocesos son:

4.1. Crear marco y seleccionar muestra

Este subproceso establece el marco y selecciona la muestra para esta iteración de la recogida, tal y como se especifica en el subproceso 2.4 (Diseñar marco y muestra). También incluye la coordinación de muestras entre repeticiones del mismo proceso estadístico (por ejemplo, para gestionar duplicidades o rotaciones) y entre diferentes procesos usando un marco o registro común (por ejemplo, para gestionar duplicidades o para distribuir la carga de respuesta). El control de calidad y la aprobación del marco y de la muestra seleccionada también tienen lugar en este subproceso, aunque el mantenimiento de registros subyacentes, de los cuales se extraen los marcos para varios procesos estadísticos, se trata como un proceso separado. Todo lo relacionado con la muestra en este subproceso no es normalmente relevante para procesos basados enteramente en el uso de fuentes preexistentes (por ejemplo, fuentes administrativas) ya que tales procesos generalmente crean marcos a partir de datos disponibles y después siguen una aproximación censal.

4.2. Inicializar recogida/obtención

Este subproceso asegura que el personal, los procesos y la tecnología están listos para recoger los datos y metadatos a través de todos los métodos según la fase de diseño. Tiene lugar a lo largo de un período de tiempo que incluye la estrategia, planificación y las actividades de formación en preparación para la ejecución específica del proceso estadístico en el caso de uso en curso. En operaciones estadísticas periódicas, algunas (o todas) de estas actividades puede que no sean necesarias de forma explícita para cada iteración. Para procesos nuevos y únicos, estas actividades pueden llevar mucho tiempo. Este subproceso incluye:

- preparar una estrategia de recogida;
- formar al personal de recogida;
- asegurar que los recursos de recogida están disponibles (por ejemplo, ordenadores portátiles o tablets);
- acordar los términos con cualquier organismo intermediario en la recogida (por ejemplo, subcontratos para las entrevistas por CATI – *Computer Assisted Telephone*

Interviewing);

- configurar los sistemas de recogida para solicitar y recibir los datos;
- asegurar la seguridad de los datos recogidos;
- preparar los instrumentos de recogida (por ejemplo, imprimir cuestionarios, prerellenarlos con los datos existentes, cargar los cuestionarios y datos en los ordenadores de los entrevistadores, etc.).

Para fuentes de datos no provenientes de encuestas, este subproceso incluirá asegurar que existen los procesos, sistemas y procedimientos de confidencialidad necesarios dentro del proceso para recibir o extraer la información necesaria de la fuente.

4.3. Ejecutar recogida/obtención

En este subproceso se implementa la recogida, junto con los diferentes instrumentos que se van a usar para recoger o recolectar la información, que puede incluir microdatos brutos o agregados producidos desde su origen, así como cualquier metadato asociado. Incluye el contacto inicial con informantes y cualquier seguimiento posterior o recordatorios. Puede incluir la entrada manual de datos durante el contacto con el informante o la gestión del trabajo de campo, dependiendo de la fuente y método de recogida. En este subproceso se graban cuándo y cómo se contacta con los informantes y si han contestado. Este subproceso también incluye la gestión de los informantes que forman parte de la recogida actual, asegurando que la relación entre la organización estadística y los informantes es positiva, grabando y respondiendo a los comentarios, peticiones y quejas. Para fuentes administrativas y otras fuentes no estadísticas, este proceso es breve: debe establecerse el contacto con el informante para que envíe la información, o la envía de acuerdo con un calendario fijado de antemano. Cuando la recogida alcanza su objetivo, se cierra y se produce un informe sobre el proceso. Algunas validaciones básicas de la estructura e integridad de la información recibida pueden tener lugar en este subproceso (por ejemplo, comprobar que los ficheros tienen el formato adecuado y contienen los campos esperados). Todas las validaciones sobre el contenido tienen lugar en la fase 'Procesar'.

4.4. Finalizar recogida/obtención

Este subproceso incluye cargar los datos y metadatos recogidos en un entorno electrónico adecuado para un procesamiento posterior. Puede incluir la recogida de datos manual o automática, por ejemplo, usando personal administrativo o herramientas de reconocimiento óptico de caracteres para extraer información de los cuestionarios en papel o convertir los formatos de ficheros recibidos de otras organizaciones. También puede incluir el análisis de los metadatos y paradatos asociados con la recogida para asegurar que tales actividades han cumplido los requisitos exigidos. En operaciones estadísticas con un instrumento físico de recogida (como un cuestionario en papel) que no se necesita para un procesamiento posterior, en este subproceso debe gestionarse el archivo de este material.

5. Procesar

Esta fase describe la depuración de los datos y su preparación para el análisis. Está formado por subprocesos que comprueban, depuran y transforman los datos de entrada de forma que puedan ser analizados y difundidos como resultados estadísticos. Se puede repetir varias veces si es necesario. Para resultados estadísticos producidos de manera periódica, esta fase tiene lugar en cada iteración. Los subprocesos en esta fase se pueden aplicar a datos tanto de fuentes estadísticas como no estadísticas (con la posible excepción del subproceso 5.6. Cálculo de pesos, que normalmente es específica de datos muestrales).

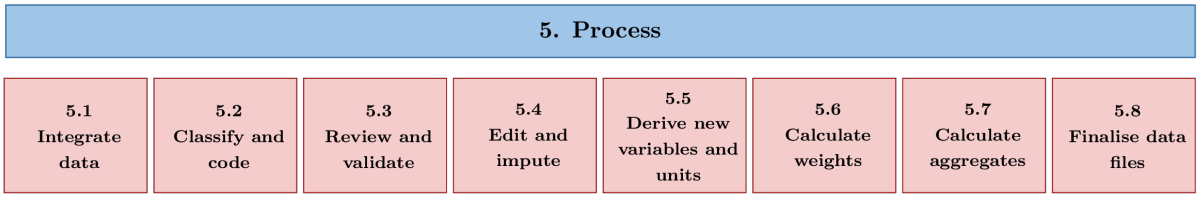

Figura 26.2: Fase 5 del GSBPM

Las fases 'Procesar' y 'Analizar' pueden ser iterativas y paralelas. El análisis (proceso 6.) puede revelar un conocimiento más amplio de los datos, que puede hacer evidente que sea necesario un procesamiento adicional en este proceso 5. Algunas actividades de las fases 'Procesar' y 'Analizar' pueden empezar antes de que la fase 'Recoger/Obtener' esté completada. Esto permite la recopilación de resultados provisionales cuando la oportunidad[2] es un requisito importante para los usuarios, incrementando así el tiempo disponible para análisis.

Esta fase se divide en ocho subprocesos, que pueden ser secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

5.1. Integrar datos

Este subproceso integra datos de una o más fuentes combinando los resultados de los subprocesos de la fase 'Recoger/Obtener'. Los datos de entrada pueden venir de una mezcla de fuentes de datos externos e internos y de una variedad de métodos de recogida, incluyendo explotaciones de datos administrativos. El resultado es un conjunto de datos enlazados (*linked data*). La integración de datos puede incluir:

- combinar datos de múltiples fuentes como parte de la creación de estadísticas integradas (como las cuentas nacionales);
- rutinas de *matching /record linkage*, con el fin de unir micro o macro datos de distintas fuentes;

[2]Traducimos *timeliness* por oportunidad.

- priorizar, cuando dos o más fuentes contienen datos para la misma variable, con valores potencialmente diferentes.

La integración de datos puede tener lugar en cualquier momento de esta fase, antes o después de cualquier otro subproceso. También puede haber varias acciones de integración de datos a lo largo del proceso estadístico. Después de la integración, dependiendo de los requisitos de la protección de datos, los datos pueden ser anonimizados, es decir, eliminando los identificadores tales como el nombre y la dirección, para ayudar a proteger la confidencialidad.

5.2. Clasificar y codificar

Este subproceso clasifica y codifica los datos de entrada. Por ejemplo, rutinas de codificación automáticas (o manuales) pueden asignar códigos numéricos a respuestas de texto de acuerdo con un esquema de clasificación predeterminado.

5.3.-4. Revisar y validar. Depurar e imputar[3]

Este subproceso examina los datos para tratar de identificar problemas potenciales, errores y discrepancias tales como valores atípicos (*outliers*), falta de respuesta parcial y codificación errónea. También se puede referir a este subproceso como validación de los datos de entrada. Debe realizarse iterativamente, validando los datos de acuerdo con unas reglas de depuración predefinidas, normalmente en un orden preestablecido. Este subproceso puede señalar/marcar datos para la inspección o depuración automática o manual. La revisión y validación se pueden aplicar a datos de cualquier tipo de fuente, antes y después de la integración. Aunque la validación es tratada en este estándar como parte de la fase 'Procesar', en la práctica algunos elementos de la validación pueden tener lugar durante la ejecución de las actividades de recogida, particularmente para casos como la recogida web o la recogida asistida por ordenador. Aunque en el estándar internacional la detección de errores reales o potenciales se enmarca en el proceso 5.3, cualquier corrección que modifique los datos tiene lugar en el subproceso 5.4. En la adaptación del INE, ambas actividades figuran juntas porque es práctica común su ejecución integrada (detección+corrección).

Cuando los datos se consideren incorrectos, faltantes (*missing*) o de poca confianza/sin consistencia, en este subproceso se pueden incluir nuevos valores. Los términos depuración e imputación abarcan una gran variedad de métodos incluyendo enfoques basados en reglas. Pasos específicos generalmente incluyen:

- la determinación sobre si incluir o cambiar datos;
- la selección del método a usar;
- añadir/cambiar los nuevos valores de datos;

[3] Aunque en la versión original del GSBPM los subprocesos 5.3 y 5.4 está separados, en el INE se ha decidido englobarlos en la adaptación del estándar ya que en la práctica, los procesos de revisión, validación, depuración e imputación se hallan altamente integrados (se ejecutan casi simultáneamente).

- escribir los nuevos valores de datos en el conjunto de datos y marcarlos como cambiados;
- la producción de metadatos sobre el proceso de depuración e imputación.

5.5. Derivar nuevas variables y unidades

Este subproceso deriva datos para variables y unidades que no se obtienen de manera explícita en la recogida, pero que son necesarias para obtener los resultados estadísticos finales. Deriva nuevas variables aplicando una fórmula matemática a una o más variables que están presentes en el conjunto de datos, o aplicando distintos supuestos de un modelo. Esta actividad puede necesitar realizarse de manera iterativa, ya que algunas variables pueden a su vez estar basadas en otras variables derivadas. Por tanto, es importante asegurar que esas variables han sido derivadas en el orden correcto. Las nuevas unidades se pueden derivar agregando o dividiendo datos de unidades de recogida o mediante otros métodos de estimación. Algunos ejemplos incluyen obtener hogares cuando la unidad de recogida son las personas o empresas cuando la unidad de recogida son unidades jurídicas.

5.6. Calcular pesos

Este subproceso crea pesos para los datos de cada unidad de acuerdo con la metodología creada en el subproceso 2.5 (Diseño del proceso y análisis). En caso de una encuesta muestral, los pesos se pueden usar para 'elevar' los resultados para hacerlos representativos de la población objetivo, o para ajustar la falta de respuesta en censos. En otras situaciones, las variables pueden necesitar pesos con fines de normalización.

En la adaptación del INE, este subproceso se concentra en el cómputo de las ponderaciones para la construcción de índices compuestos a partir de índices simples para todos los grados de desagregación posible considerados en el diseño de la operación. El cómputo de pesos de muestreo calibrados, por el contrario, se considera una actividad altamente integrada con la construcción de estimadores y cálculo de agregados, por ello aparece en el proceso 5.7. Calcular agregados.

5.7. Calcular agregados

Este subproceso crea datos agregados y totales poblacionales a partir de microdatos o de agregados de nivel inferior. Incluye sumar datos de registros que comparten determinadas características, calculando medidas de promedios (media, mediana) y dispersión y aplicando los pesos del subproceso 5.6 para calcular índices compuestos adecuados. En el caso de una operación muestral, los errores muestrales (que involucra la estimación de varianzas) también se calculan en este subproceso asociándolos a los agregados correspondientes.

En la adaptación del INE, para operaciones estadísticas muestrales este subproceso incluye el cálculo de pesos de muestreo calibrados.

5.8. Finalizar ficheros de datos

Este subproceso integra los resultados de los otros subprocesos de esta fase y los resultados en un fichero de datos (normalmente de macrodatos), que se utiliza como el input de la fase 'Analizar'. Algunas veces puede ser un fichero intermedio más que uno final, particularmente para procesos de negocios en los que hay fuertes restricciones de tiempo y una necesidad de producir estimadores tanto preliminares como finales.

6. Analizar

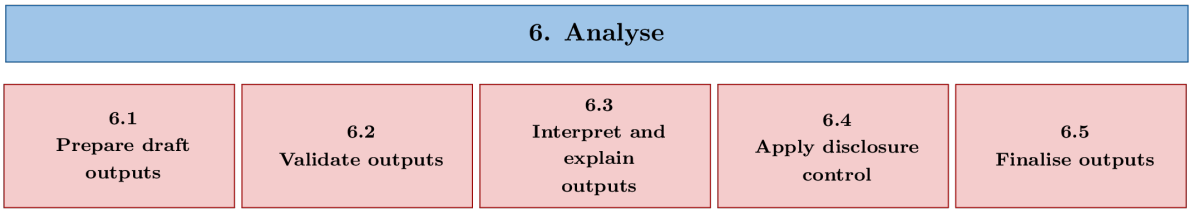

Figura 26.3: Fase 6 del GSBPM

En esta fase se producen resultados estadísticos que se examinan al detalle y se preparan para su difusión. Incluye preparar el contenido estadístico (incluyendo comentarios, notas técnicas, etc.), y asegurar que los resultados son 'adecuados para su propósito' antes de la difusión a los usuarios. Esta fase también incluye los subprocesos y actividades que permiten a los analistas estadísticos entender las estadísticas producidas. Para resultados estadísticos producidos periódicamente, esta fase ocurre en cada iteración. La fase 'Analizar' y sus subprocesos son genéricos para todos los resultados estadísticos, independientemente de la fuente que se haya utilizado.

La fase 'Analizar' se divide en cinco subprocesos, que pueden ser secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

6.1. Preparar borrador de resultados

En este subproceso se transforman los datos en resultados estadísticos. Incluye la producción de outputs como índices, tendencias o series ajustadas de estacionalidad, así como la documentación de los indicadores de calidad.

6.2.-3. Validar resultados. Interpretar y explicar los resultados[4]

En este subproceso los estadísticos evalúan y validan la calidad de los resultados producidos de acuerdo con el marco general de calidad y con las expectativas. Este

[4] Aunque en la versión original del GSBPM los subprocesos 6.2 y 6.3 está separados, en el INE se ha decidido englobarlos ya que en la práctica, los procesos de validación, interpretación y explicación de resultados están altamente integrados.

subproceso también incluye actividades relacionadas con la recogida de información, con el efecto a largo plazo de construir un conjunto acumulado de conocimientos sobre un dominio estadístico específico. Este conocimiento se aplica a la recogida concreta, en el entorno concreto de producción en curso, para identificar cualquier divergencia de las expectativas y hacer posibles análisis bien fundamentados. Las actividades de validación pueden incluir:

- comprobar que la cobertura de la población y las tasas de respuesta son las requeridas;
- comparar los resultados con los ciclos previos (si es pertinente);
- comprobar que se dispone de los metadatos y paradatos (valores que adquieren los metadatos cuando se ejecuta el proceso de producción) asociados y que están en línea con las expectativas;
- confrontar los resultados con otros datos relevantes (tanto internos como externos);
- investigar las inconsistencias en los estadísticos, agregados, índices y tasas de variación;
- llevar a cabo la depuración macro;
- validar los estadísticos, agregados, índices y tasas con las expectativas y el dominio de conocimiento.

En este subproceso tiene lugar el entendimiento y comprensión en profundidad de los resultados por parte de los estadísticos. Los expertos usan este conocimiento para interpretar y explicar las estadísticas producidas para este ciclo evaluando en qué medida las estadísticas reflejan las expectativas iniciales, observando las estadísticas desde todas las perspectivas usando distintas herramientas y medios y llevando a cabo profundos análisis estadísticos.

6.4. Aplicar control del secreto estadístico

Este subproceso asegura que los datos (y metadatos) que se van a publicar no violan las reglas y normas legales de confidencialidad. Esto debe controlarse para comprobar el control del secreto estadístico primario y secundario, así como la aplicación de supresión de datos o las técnicas de perturbación. El grado y método de anonimización del control del secreto estadístico puede variar para distintos tipos de resultados, por ejemplo, el enfoque usado para conjuntos de microdatos con fines de investigación será diferente al que se usa para publicar tablas o mapas.

6.5. Finalizar resultados

Este subproceso asegura que las estadísticas y la información asociada son adecuados para el propósito y alcanzan el nivel de calidad requerido y, por tanto, están preparados para su uso. Incluye:

- completar los controles de consistencia;

- determinar el nivel de difusión y aplicar excepciones;
- recopilar información complementaria, incluyendo interpretación, comentarios, notas técnicas, instrucciones/resúmenes, medidas de incertidumbre y cualquier otro metadato necesario;
- producir la documentación adicional interna;
- discutir previamente a la difusión con adecuados expertos internos en la materia;
- traducir los resultados a varios idiomas en países multilingües;
- aprobar el contenido estadístico para la difusión.

7. Difundir

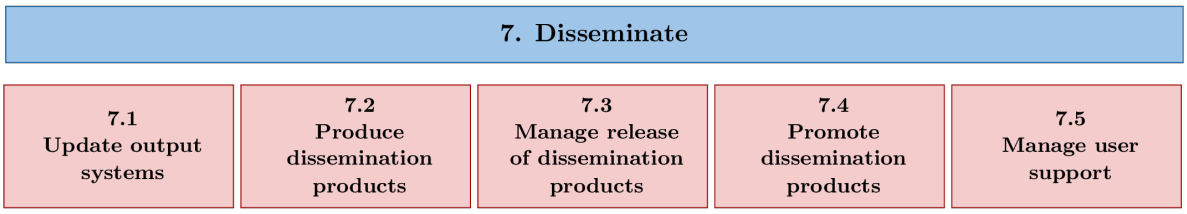

Figura 26.4: Fase 7 del GSBPM

Esta fase gestiona la difusión de los productos estadísticos para los usuarios. Incluye todas las actividades asociadas a la recopilación y publicación de un conjunto de productos estáticos y dinámicos vía una variedad de canales. Estas actividades ayudan al usuario en el acceso y uso de los resultados publicados por la organización estadística.

Para resultados estadísticos producidos de manera periódica, esta fase ocurre en cada iteración. Está formada por cinco subprocesos, que pueden ser secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

7.1. Actualizar sistemas de resultado

Este subproceso gestiona la actualización de los sistemas en los que los datos y metadatos se almacenan cuando están listos para su difusión, incluyendo:

- formatear datos y metadatos que están listos para poner en las bases de datos de resultados;
- cargar los datos y metadatos en las bases de datos de resultados;
- asegurar que los datos están enlazados con los metadatos correspondientes.

El formateo, la carga y el enlace de metadatos debe llevarse a cabo preferiblemente en fases anteriores, pero este subproceso incluye una revisión final de que todos los

metadatos necesarios están en su sitio listos para la difusión.

7.2. Producir productos de difusión

Este subproceso produce los productos según el diseño previo (en el subproceso 2.1) para alcanzar las necesidades de los usuarios. Puede incluir publicaciones impresas, notas de prensa y páginas web. Los productos pueden tener muchas formas incluyendo gráficos interactivos, tablas, conjuntos de microdatos de uso público y ficheros que se pueden descargar. Los pasos típicos incluyen:

- preparar las componentes del producto (texto explicativo, tablas, gráficos, indicadores de calidad, etc.);
- juntar los componentes en productos;
- editar los productos y comprobar que alcanzan los estándares de publicación.

7.3. Gestionar divulgación de productos de difusión

Este subproceso asegura que existen todos los elementos para la difusión incluyendo la gestión del momento adecuado (*timing*) de la difusión. Incluye la difusión de la información para grupos específicos como la prensa o ministerios, así como acuerdos para embargos previos a la difusión. También incluye la provisión de productos a subscriptores y gestionar el acceso a datos confidenciales por parte de grupos de usuarios autorizados, como investigadores. Algunas veces una organización puede necesitar retractar un producto, por ejemplo, si se descubre un error. Esto también se incluye en este subproceso.

7.4. Promocionar productos de difusión

Mientras que el marketing en general se puede considerar un proceso global (*overarching process*), este subproceso está relacionado con una promoción activa de los productos estadísticos producidos en un proceso de negocio estadístico específico, de modo que se procure alcanzar la mayor audiencia posible. Incluye el uso de herramientas que gestionan la relación con los clientes, para dirigirse mejor a los potenciales usuarios de los productos, así como el uso de herramientas incluyendo páginas web, wikis y blogs para facilitar el proceso de comunicar la información estadística a los usuarios.

7.5. Gestionar soporte al usuario

Este subproceso asegura que las consultas y peticiones de servicios de los usuarios tales como el acceso a los microdatos se registran y procesan y que las respuestas se proporcionan dentro de las fechas límites acordadas. Estas consultas y peticiones deben ser revisadas de forma periódica para proporcionar un input al proceso global (*overarching process*) de control de calidad, ya que pueden indicar necesidades nuevas o cambiantes de los usuarios.

8. Evaluar

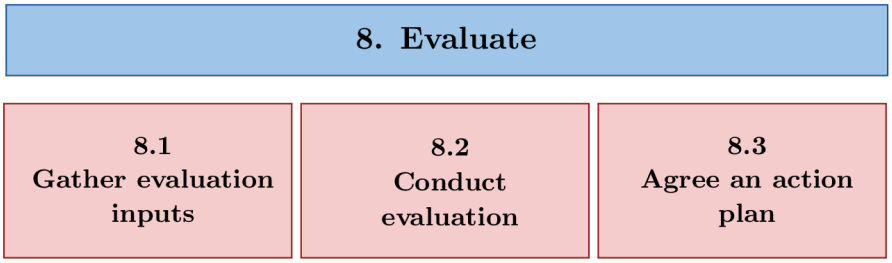

Figura 26.5: Fase 8 del GSBPM

Esta fase gestiona la evaluación de una ejecución específica del proceso, en contraposición con el proceso más general de gestión de la calidad estadística descrita en la sección siguiente. Lógicamente tiene lugar al final de la ejecución del proceso, pero depende de los inputs recogidos a lo largo de las diferentes fases. Incluye la evaluación del éxito de una ejecución específica del proceso de negocio estadístico, haciendo uso de un conjunto de inputs cuantitativos y cualitativos e identificando y priorizando mejoras potenciales.

Para resultados estadísticos producidos de manera periódica, la evaluación debe, por lo menos en teoría, tener lugar en cada iteración, determinando si futuras iteraciones deberían tener lugar, y en caso de que sea así, si debe implementarse alguna mejora. Sin embargo, en algunos casos, particularmente para procesos de negocio estadísticos periódicos bien consolidados, la evaluación puede no llevarse a cabo de manera formal en cada iteración. En tales casos, esta fase se puede ver como la forma de decidir si la próxima iteración debería empezar desde la fase 'Especificar necesidades' o desde algún fase posterior (a menudo la fase 'Recoger/Obtener').

Esta fase está compuesta por tres subprocesos, que generalmente son secuenciales, de izquierda a derecha, pero que se pueden solapar en la práctica. Estos subprocesos son:

8.1. Reunir inputs para la evaluación

El material para la evaluación puede producirse en cualquier otra fase o subproceso. Puede tomar muchas formas, incluyendo *feedback* de los usuarios, metadatos de proceso, métricas del sistema y sugerencias del personal. Los informes sobre el progreso de un plan de acción acordado en una iteración previa también pueden ser un input para evaluaciones de iteraciones posteriores. Este subproceso reúne todos estos inputs y los hace disponibles para la persona o equipo que lleve a cabo la evaluación.

8.2. Ejecutar evaluación

Este subproceso analiza los inputs de evaluación y los sintetiza en un informe de evaluación. El informe resultante debe tomar nota de cualquier asunto de calidad específico

de esta iteración del proceso de negocio estadístico y debe hacer recomendaciones sobre cambios en caso de sea apropiado. Estas recomendaciones pueden abarcar cambios de cualquier fase o subproceso para futuras iteraciones del proceso o pueden sugerir que el proceso no se repita.

8.3. Acordar un plan de acción

Este subproceso ejecuta la toma de decisión necesaria para dar forma y acordar un plan de acción basado en el informe de evaluación. También debe de incluir consideraciones sobre un mecanismo para la supervisión del impacto de esas acciones, que pueden, a su vez, proporcionar un input para evaluaciones de futuras iteraciones del proceso.

26.2 Procesos generales (*overarching processes*)

El GSBPM también reconoce varios procesos generales que se aplican durante todas las fases de producción y a través de los procesos. Algunos de estos procesos generales se mencionan en el tema 25. Los procesos de gestión de la calidad y gestión de los datos y metadatos se explican más detenidamente en esta sección.

Gestión de la Calidad

La calidad afecta a las organizaciones, procesos y productos. En el marco actual, el proceso general de gestión de la calidad se refiere a la calidad del producto y del proceso. La calidad a nivel institucional (por ejemplo, la adopción de una Política de Calidad o de un Marco de Garantía de Calidad) está incluido en el GAMSO (UNECE 2019a).

El objetivo de la gestión de la calidad dentro del proceso estadístico es entender y gestionar la calidad de los productos estadísticos. Existe un acuerdo general entre las organizaciones estadísticas de que la calidad debería estar definida de acuerdo con la norma ISO 9000-2005 (ISO9000:2005 2005): 'Grado en el que un conjunto de características inherentes cumplen con los requisitos'. Por tanto, la calidad del producto es un concepto complejo y multifacético, normalmente definido en términos de varias dimensiones de calidad. Las dimensiones de la calidad que se consideran más importantes dependen de las perspectivas del usuario, de sus necesidades y prioridades, que varían entre procesos y grupos de usuarios.

Con el fin de mejorar la calidad del producto, la gestión de la calidad debe estar presente a lo largo de todo el modelo del proceso estadístico. Existe una estrecha relación con la Fase 8 (Evaluar), que tiene el papel específico de evaluar a posteriori casos y ejecuciones individuales de procesos estadísticos. Sin embargo, la gestión de la calidad tiene una cobertura más profunda y más amplia. Así como la evaluación de las iteraciones de un proceso, también es necesario evaluar de forma separada las fases y subprocesos, lo óptimo sería cada vez que se utilizan, pero por lo menos de acuerdo con un calendario

acordado. Los metadatos generados por los distintos subprocesos en sí mismos también resultan de interés como input para la gestión de la calidad del proceso. Estas evaluaciones se pueden implementar en procesos específicos o entre varios procesos que usen componentes comunes.

Además, el conjunto de acciones que deben implementarse en los subprocesos para evitar y controlar los errores juegan un papel fundamental en la gestión de la calidad. La estrategia se puede incluir en el plan de garantía de la calidad.

Dentro de una organización, la gestión de la calidad se referirá generalmente a un marco específico de calidad y, por tanto, puede tomar distintas formas y ofrecer distintos resultados en diferentes organizaciones. La multiplicidad de marcos de calidad existentes aumenta la importancia de los estudios comparativos y las evaluaciones mediante revisión por pares (*peer review*) y aunque resulta dudoso que estos enfoques sean factibles para cada iteración de cada parte de cada proceso estadístico, debe usarse de forma sistemática de acuerdo con un calendario acordado que permita la revisión de las principales partes del proceso en un período de tiempo específico.

Ampliando el campo de aplicación del proceso general de gestión de la calidad, también se puede considerar la evaluación de grupos de procesos estadísticos con el fin de identificar potenciales duplicaciones o lagunas.

Todas las evaluaciones darán lugar a *feedback*, que debe usarse para mejorar el proceso, fase o subproceso relevante, creando un bucle de calidad.

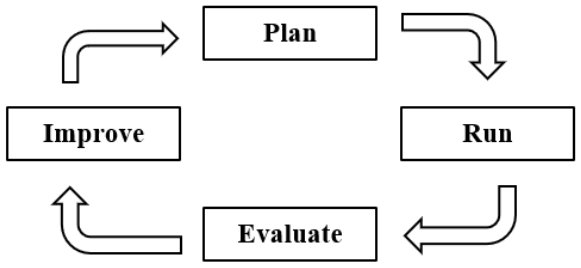

Figura 26.6: Bucle de calidad

Ejemplos de actividades de gestión de la calidad incluyen:

- Establecer y mantener el marco de calidad;
- Establecer criterios globales de calidad;
- Establecer los objetivos de calidad del proceso y supervisar su cumplimiento;
- Solicitar y analizar el *feedback* de los usuarios;
- Revisar las acciones y la documentación de las lecciones aprendidas;

- Examinar los metadatos de proceso y los indicadores de calidad;
- Auditorías internas o externas de procesos estadísticos.

Los indicadores de calidad sirven de ayuda a la gestión de calidad orientada a los procesos. Una lista de indicadores de calidad para las fases y subfases del GSBPM así como para los procesos generales de gestión de la calidad y de los metadatos se puede encontrar en los Indicadores de Calidad del GSBPM (UNECE 2018). Entre otros, se pueden usar para identificar carencias y/o duplicidades de trabajo en la organización.

Gestión de los metadatos

Los metadatos juegan un papel muy importante y deben ser gestionados a un nivel operativo dentro del proceso de producción estadística. Cuando los aspectos de la gestión de los metadatos se consideran a nivel corporativo o estratégico (p.ej. hay sistemas de metadatos que afectan a grandes partes del sistema de producción) deberían ser considerados en el GAMSO (UNECE 2019a).

Una buena gestión de los metadatos es esencial para un eficiente funcionamiento del proceso estadístico. Los metadatos están presentes en cada fase, tanto si son creados como si son transferidos de una fase previa. En el contexto de este modelo, el énfasis del proceso general de gestión de los metadatos está en la creación, uso y archivo de metadatos estadísticos, aunque los metadatos de los distintos subprocesos en sí mismos también son de interés, incluyéndolos como un input para la gestión de calidad. El principal problema es asegurar que estos metadatos son recogidos tan pronto como sea posible y almacenados y transferidos de una fase a otra junto con los datos a los que se refieren. La estrategia y los sistemas de gestión de los metadatos son, por tanto, vitales para el funcionamiento de este modelo y pueden ser facilitados por el GSIM (UNECE 2019b).

El GSIM es un marco de referencia de objetos de información que permite descripciones genéricas de la definición, gestión y uso de datos y metadatos durante todo el proceso de producción. El GSIM favorece un enfoque consistente para los metadatos, facilitando el papel básico de los metadatos, es decir, que estos deben definir de forma única y formalmente el contenido y los enlaces entre los objetos de información y los procesos de producción en el sistema de información estadística.

El Marco Común de Metadatos METIS (*METIS Common Metadata Framework*) (UNECE 2021b) identifica los siguientes dieciséis principios esenciales para la gestión de los metadatos, todos ellos previstos que se cubran en el proceso general de gestión de los metadatos y tenidos en consideración cuando se diseñe e implemente el sistema de metadatos estadísticos. Los principios se presentan en cuatro grupos:

Tratamiento de los metadatos	
	i. Modelo del Proceso de Negocio Estadístico: Gestiónense los metadatos con atención al conjunto del modelo de proceso estadístico. ii. Activo no pasivo: Háganse los metadatos activos en la mayor medida posible. Metadatos activos son metadatos que conducen a otros procesos y acciones. Tratando los metadatos de esta forma se asegurará que son acurados y actualizados. iii. Reutilización: Reutilícense los metadatos donde sea posible para la integración estadística así como por razones de eficiencia. iv. Versiones: Presérvese la historia (versiones anteriores) de los metadatos.
Autoridad de metadatos	
	i. Registro: Asegúrese que el proceso de registro (flujo de trabajo) asociado con cada elemento de los metadatos está bien documentado de forma que esté clara la identificación del titular, estado de aprobación, fecha de la operación, etc. ii. Fuente única: Asegúrese que existe una única fuente, autorizada ('autoridad de registro´) para cada elemento de los metadatos. iii. Una entrada/actualización: Minimícense los errores con una entrada de datos única y realizando actualizaciones en un único sitio. iv. Variaciones de los estándares: Asegúrese que las variaciones de los estándares son estrictamente gestionadas/aprobadas, documentadas y visibles.
Relación con el Ciclo/Proceso Estadístico	
	i. Integridad: Hágase del trabajo relacionado con los metadatos una parte integral de los procesos en toda la organización. ii. Enlace de metadatos: Asegúrese que los metadatos presentados a los usuarios enlazan con los metadatos que condujeron el proceso o que se crearon durante el proceso. iii. Describir el flujo: Descríbase el flujo de metadatos con los procesos de negocios y estadístico (junto con el flujo de datos y la lógica de negocio). iv. Captura en la fuente: Captúrense los metadatos en su fuente, preferiblemente de forma automática como un subproducto de otros procesos. v. Intercambio y uso: Intercámbiense los metadatos y úsense para proporcionar información tanto para procesos informatizados automáticos como para la interpretación humana. La infraestructura para intercambio de datos y sus metadatos asociados debería estar basada en componentes sin conexión directa, con la opción de lenguajes estándares de intercambio, como el XML.
Usuarios	
	i. Identifica usuarios: Asegúrese que los usuarios se identifican con claridad para todos los procesos relativos a los metadatos y que todos los metadatos capturados generarán valor para ellos. ii. Distintos formatos: La diversidad de metadatos debe reconocerse de modo que haya distintos puntos de vista correspondientes con los distintos usos de los datos. Los distintos usuarios requieren distintos niveles de detalle. Los metadatos aparecen en diferentes formatos dependiendo de los procesos y objetivos para los que se han producido y usado. iii. Disponibilidad: Asegúrese que los metadatos están disponibles pronto y aprovechables en el contexto de las necesidades de información de los usuarios (tanto si el usuario es interno o externo).

Gestión de los datos

La gestión de datos es esencial ya que los datos se producen en muchas de las activida-

des de los procesos y son los principales outputs. El principal objetivo de la gestión de los datos es asegurar que los datos se usan de forma apropiada y son útiles a lo largo de su ciclo de vida. La gestión de los datos a lo largo de su ciclo de vida incluye actividades como la planificación y la evaluación de los procesos de gestión de los datos así como la creación y la implementación de procesos relacionados con la recogida, organización, uso, protección, preservación y eliminación de los datos.

Cómo los datos son gestionados estará muy vinculado al uso de los datos, que por otro lado está relacionado con los procesos estadísticos donde los datos son creados. Tanto los datos como los procesos en los que son creados deben estar bien definidos con el fin de asegurar una gestión adecuada de los datos.

Ejemplos de actividades de gestión de los datos incluyen:

- Establecer una estructura de gobierno y asignar responsabilidades administrativas de los datos;
- Diseñar estructuras de datos y sus conjuntos de datos asociados y el flujo de datos a lo largo del proceso estadístico;
- Identificar bases de datos (repositorios) para almacenar los datos y la gestión de la base de datos;
- Documentar los datos (p.ej. registrando e inventariando los datos, clasificando los datos de acuerdo con su contenido, permanencia u otras clasificaciones necesarias);
- Determinar los periodos de conservación de los datos[5];
- Asegurar los datos frente al acceso y uso no autorizado;
- Proteger los datos frente a cambios tecnológicos, deterioro de los medios físicos y corrupción de datos;
- Realizar comprobaciones de la integridad de los datos (p.ej. comprobaciones periódicas que aseguren la acuracidad y consistencia de los datos a lo largo de su ciclo de vida);
- Realizar actividades de tratamiento una vez que el periodo de conservación de los datos ha expirado.

26.3 Otros usos del GSBPM

El objetivo original del GSBPM era proporcionar una base para que las organizaciones estadísticas acordaran una terminología estándar para ayudar en sus discusiones sobre el desarrollo de un sistema de metadatos estadísticos y de procesos. Sin embargo, a medida que el modelo se ha desarrollado, ha resultado cada vez más aparente que se

[5]Por conservación (*retention*) definimos las políticas de gestión de datos y registros permanentes para cumplir los requisitos legales y de negocio de archivación de datos (Wikipedia 2021).

puede usar para muchos otros propósitos, en particular, relacionados con la modernización de estadísticas oficiales. Los artículos y documentos que describen los usos actuales y potenciales del GSBPM están disponibles en la wiki de la UNECE (UNECE 2021a). La lista que figura a continuación tiene por objetivo señalar algunos usos actuales e inspirar nuevas ideas sobre cómo el GSBPM se puede usar en la práctica.

- Proporcionar una estructura para la documentación de procesos estadísticos.- El GSBPM puede proporcionar una estructura para organizar y acumular la documentación dentro de una organización, promocionando la estandarización y la identificación de buenas prácticas;
- Facilitar la compartición de los métodos estadísticos y el software.- El GSBPM denie los componentes de los procesos estadísticos de modo que no solo fomenta el intercambio de métodos y herramientas informáticas, sino que también facilita el intercambio entre distintas organizaciones estadísticas que utilizan el modelo;
- Describir los estándares que son utilizados o podrían ser usados para las distintas fases del proceso estadístico. Por ejemplo, el Anexo 2 de la guía de usuario del estándar SDMX 2.1 (SDMX 2012) explora cómo el SDMX se usa en el trabajo estadístico en el contexto de un modelo de procesos de negocio;
- Proporcionar un marco para la evaluación y mejora de la calidad del proceso.– Si un enfoque por comparación a la evaluación de la calidad del proceso resulta exitoso, es necesario estandarizar procesos tanto como sea posible. El GSBPM proporciona un mecanismo para facilitar esto;
- Integrar mejor los trabajos sobre metadatos y calidad estadísticos.– Enlazado con el punto anterior, el marco común proporcionado por el GSBPM puede ayudar a integrar el trabajo internacional sobre los metadatos estadísticos con aquél sobre calidad de datos proporcionando un marco y una terminología comunes para describir el proceso estadístico;
- Proporcionar el modelo básico para marcos estándares metodológicos.– Los estándares metodológicos se pueden enlazar con la(s) fase(s) o subproceso(s) con los que están relacionados y pueden entonces ser clasificados y almacenados en una estructura basada en el GSBPM;
- Desarrollar un modelo de repositorio de proceso de negocio para almacenar outputs del modelado de procesos y enlazarlos con el modelo de proceso de negocio estadístico;
- Proporcionar un marco subyacente para desarrollar y un conjunto de terminología estándar para describir competencias y *expertise* necesarios en el proceso de producción estadística;
- Medir los costes operacionales.- El GSBPM se puede usar como una base para medir el coste de las distintas partes de un proceso estadístico. Esto ayuda a localizar actividades de modernización que mejoren la eficiencia de las partes del proceso que son más costosas;
- Medir el rendimiento del sistema.– Relacionado con el punto anterior sobre los

costes, el GSBPM también se puede usar para identificar componentes que no se están realizando de forma eficiente, que se están duplicando una a otra innecesariamente o que necesitan ser reemplazadas. De forma similar pueden identificarse carencias para las cuales se deben desarrollar nuevos componentes;

- Proporcionar una herramienta para alinear procesos de negocio de proveedores de datos no estadísticos (p.ej. datos administrativos o geoespaciales) facilitando la comunicación entre estadísticos y expertos de otros dominios y para armonizar la terminología relacionada;
- Proporcionar una herramienta para fortalecer la capacidad y el conocimiento técnico metódicamente mediante referencias detalladas a cada fase de producción;
- Proporcionar una herramientas para el desarrollo y revisión de las clasificaciones estadísticas.

26.4 Data Documentation Initiative (DDI), SDMX y comparación con el GSBPM

El modelo combinado del ciclo de vida DDI 3

Este modelo se ha desarrollado en el consorcio *Data Documentation Initiative* (DDI) (DDI Alliance 2021), una iniciativa internacional para establecer un estándar de documentación técnica para describir los datos en ciencias sociales. La Alianza DDI incluye principalmente instituciones académicas y de investigación, por tanto, el alcance de este modelo es un poco distinto del GSBPM, que se aplica de manera específica a las organizaciones de estadística oficial. A pesar de esto, el proceso estadístico parece bastante similar entre los productores de estadísticas oficiales y no oficiales, como puede observarse en la consistencia a alto nivel entre los modelos.

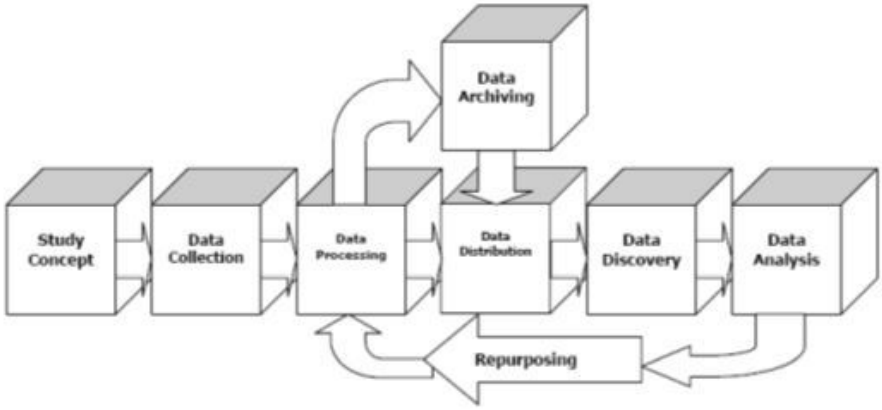

Figura 26.7: El modelo combinado del ciclo de vida DDI 3

Las principales diferencias entre el GSBPM y el modelo combinado del ciclo de vida DDI son:

- El GSBPM generalmente asume que el proceso general es llevado a cabo por una única organización (aunque para procesos grandes como los censos algunos

subprocesos pueden externalizarse). El modelo DDI parece reconocer que algunos pases como 'Análisis de datos' y 'Reutilización' pueden ser llevados a cabo por organizaciones distintas de la que recoge los datos. Esto refleja una diferencia fundamental entre las prácticas en las comunidades investigadora y de estadísticas oficiales, donde la comunidad investigadora tiene mayores posibilidades de colaboración entre organizaciones durante el proceso de producción.

- El modelo DDI reemplaza la fase de difusión por 'Distribución de Datos' que tiene ligar antes de la fase de análisis. Esto refleja la diferencia en enfoque entre las comunidades investigadora y de estadísticas oficiales, donde la última pone un mayor énfasis en la difusión de los datos, más que en una investigación basada en la difusión por parte de terceros.
- El modelo DDI contiene el proceso de 'Reutilización', definido como el uso secundario de un conjunto de datos o la creación de un conjunto de datos real o virtual armonizado. Esto generalmente se refiere a algún reuso de un conjunto de datos que no fue originalmente previsto en las fases de diseño y recogida. En el GSBPM, si los outputs de algún proceso son reutilizados para algún otro propósito, se tratan en dos procesos separados (dos casos del modelo). El segundo proceso identifica los datos en la fase 1 (Especificar Necesidades) donde hay un subproceso para confirmar la disponibilidad de datos existentes, y los obtiene en la fase 4 (Recoger/Obtener) y luego los usa para producir nuevos outputs.
- El modelo DDI tiene fases separadas para la localización de los datos y el análisis de datos, mientras que en el GSBPM estas funciones están combinadas en la fase 6 (Analizar). En algunos casos, los elementos de la fase de análisis del GSBPM también pueden estar incluidas en la fase 'Procesamiento de Datos' del DDI, dependiendo de la extensión del trabajo analítico anterior a la fase 'Distribución de Datos'.
- El GSBPM identifica de forma explícita 'procesos generales', como la calidad y la gestión de metadatos, mientras que éstos son más implícitos en el modelo DDI.

Está claro que el GSBPM y el modelo combinado del ciclo de vida DDI sirven para propósitos ligeramente diferentes. Esto proporciona una buena justificación para las diferencias entre ellos. Sin embargo, a pesar de estas diferencias en su finalidad también hay un montón de similitudes. Por tanto, es útil establecer una correspondencia entre los dos modelos para intentar tener un mejor conocimiento de cómo interactúan.

SDMX

El estándar SDMX (*Statistical Data and Metadata eXchange*) (SDMX 2012) no proporciona un modelo para procesos estadísticos en el mismo sentido que el GSBPM. Pero proporciona una terminología estándar para datos y metadatos estadísticos, así como estándares técnicos y directrices orientadas hacia el contenido para la transferencia de datos y metadatos, que se podrían aplicar entre subprocesos dentro de una organización estadística. Por tanto, se considera que se puede incorporar el GSBPM en las directrices orientadas hacia el contenido del SDMX como un dominio trnasversal. También se

considera que el SDMX puede proporcionar el formato para la transmisión de datos entre subprocesos dentro de una organización estadística.

Bibliografía

DDI Alliance (2021). *Data Documentation Initiative*. URL: https://ddialliance.org/.

ISO9000:2005 (2005). *Sistemas de gestión de la calidad*. URL: https://www.iso.org/obp/ui/es/#iso:std:iso:9000:ed-4:v1:es.

SDMX (2012). *SDMX 2.1 User Guide*. URL: https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf.

UNECE (2018). *Quality Indicators for the GSBPM*. URL: https://statswiki.unece.org/display/GSBPM/Quality+Indicators.

– (2019a). *Generic Activity Model for Statistical Organizations*. URL: https://statswiki.unece.org/display/GAMSO/.

– (2019b). *Generic Statistical Information Model v1.2*. URL: https://statswiki.unece.org/display/gsim/.

– (2019c). *The Generic Statistical Business Process Model v5.1*. URL: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

– (2021a). *Statistics Wiki*. URL: https://statswiki.unece.org/.

– (2021b). *The Common Metadata Framework*. URL: https://statswiki.unece.org/display/hlgbas/The+Common+Metadata+Framework.

Wikipedia (2021). *Data Retention*. Página visitada el día 15 de septiembre de 2021. URL: https://en.wikipedia.org/wiki/Data_retention.

Tema 27

Introducción a Data Mining. Introducción a la exploración de datos. Visión básica de estrategias de clasificación, árboles de decisión. Conceptos básicos de análisis cluster.

Este tema está elaborado usando la siguiente bibliografía.

G. James y col. (2013). *An introduction to statistical learning*. Vol. 112. Springer

Trevor Hastie, Robert Tibshirani y Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media

Edwin De Jonge y Mark Van Der Loo (2013). *An introduction to data cleaning with R*. Statistics Netherlands Heerlen

27.1 Introducción a *Data Mining*

El aprendizaje estadístico (*machine learning*) se refiere a un conjunto de herramientas para modelar y comprender conjuntos de datos complejos. Es un área de estadística desarrollada recientemente y se combina con desarrollos paralelos en informática y, en particular, sobre aprendizaje automático. El campo abarca muchos métodos, como la regresión Sparse y de Lasso, los árboles de clasificación y regresión, y las máquinas de vectores de soporte (SVM del inglés *Support Vector Machines*).

Con la explosión de problemas de 'Big Data', el aprendizaje estadístico se ha convertido en un campo muy candente en muchas áreas científicas, así como en marketing, finanzas y otras disciplinas comerciales.

Además de lo anteriormente indicado, hay que decir que el campo de la Estadística se ve en constante desafía por los problemas planteados en la ciencia y la industria. Al principio, estos problemas a menudo provenían de experimentos agrícolas e industriales y tenían un alcance relativamente pequeño. Con la aparición de los ordenadores y la era

de la información, los problemas estadísticos se han disparado tanto en tamaño como en complejidad. Los desafíos en las áreas de almacenamiento, organización y búsqueda de datos han llevado al nuevo campo de la 'minería de datos' (*Data Mining*). Los problemas estadísticos y computacionales en biología y medicina han creado la 'bioinformática'. Se están generando grandes cantidades de datos en muchos campos, y el trabajo del estadístico es darle sentido a todo: extraer patrones y tendencias importantes y comprender 'lo que dicen los datos'. A esto lo llamamos aprender de los datos y los desafíos para aprender de los datos han llevado a una revolución en las ciencias estadísticas. Dado que la computación juega un papel tan importante, no es sorprendente que gran parte de este nuevo desarrollo haya sido realizado por investigadores en otros campos como la informática y la ingeniería. Los problemas de aprendizaje que consideramos pueden clasificarse aproximadamente como supervisados o no supervisados. En el aprendizaje supervisado, el objetivo es predecir el valor de una medida de resultado basada en una serie de medidas de entrada; en el aprendizaje no supervisado, no hay una medida de resultado y el objetivo es describir las asociaciones y patrones entre un conjunto de medidas de entrada.

A partir de aquí, los siguientes apartados se centrarán en la exploración de datos, con el objetivo de partir de unos datos consistentes y confiables; la visión estratégica de la clasificación de datos, focalizándose en los árboles de decisión; y por último, los conceptos básicos del análisis cluster, como complemento a otras técnicas.

27.2 Introducción a la exploración de datos

La calidad de los resultados estadísticos depende en gran medida de la calidad de los datos subyacentes que se han tomado de base. Dado que los datos brutos a menudo son inconsistentes o incompletos, la depuración de datos puede consumir una cantidad sustancial de los recursos disponibles para análisis estadísticos. Aunque muchos programas estadísticos tiene una considerable cantidad de características para analizar datos, la funcionalidad de verificación de datos y localización de errores basada en edits es actualmente limitada. Por ello, existen en el mercado paquetes informáticos específicos que están diseñados y orientados para ofrecer una *tool box* fácil de usar para editar la definición y la manipulación, verificar los datos, y localizar los errores.

Hace unos años, las versiones anteriores de estos paquete podían manejar conjuntos de datos numéricos o categóricos. Si bien, ahora disponen de nuevas funcionalidades relacionadas con datos numéricos y condicionales mixtos, así como funcionalidades relacionadas con las restricciones condicionales, posibilidad de leer ediciones de archivos de texto de formato libre y localizar los errores más rápidamente bajo ciertas condiciones.

La depuración de datos es muy complicada por el hecho de que las reglas de depuración a menudo están interrelacionadas: una variable puede estar incluida en más de un edit y un edit puede contener múltiples variables. Por ejemplo, una variable en el saldo de una

cuenta puede estar incluida en varias reglas de suma, creando una dependencia entre esas reglas. Es común en la literatura de depuración de datos distinguir entre restricciones lineales, restricciones categóricas y restricciones condicionales. Las restricciones lineales son restricciones lineales que pueden ser no iguales, como restricciones de rango y reglas de suma que pertenecen a datos numéricos. Las restricciones categóricas son reglas que excluyen combinaciones de valores no válidos de un conjunto de datos categórico.

Por tanto, si bien la exploración de datos se da sobre un conjunto de datos muestrales o poblaciones, estos han de estar depurados y ser consistentes. Lo que nos lleva a pensar que existe una importante etapa previa de depuración, o de exploración de posibles inconsistencias o falta de datos.

Según De Jonge y Van Der Loo 2013, un conjunto de datos es una colección de datos que describe valores de variables de varias unidades del mundo real. Con datos que son técnicamente correctos, entendemos un conjunto de datos donde cada valor:

1. Puede reconocerse directamente como perteneciente a una determinada variable;
2. Se almacena en un tipo de datos que representa el dominio en el que se mueve el valor de la variable del mundo real.

En otras palabras, para cada unidad, una variable de texto debe almacenarse como texto, una variable numérica como un número, y así sucesivamente, y todo esto en un formato que sea consistente en todo el conjunto de datos. Entendemos por consistentes los datos o variables que: (a) están dentro de su rango o dominio (p.e. la altura debe ser siempre mayor que 0); (b) verifican restricciones cruzadas entre dos o más variables pudiendo existir correlación o no correlación entre ellas (p.e., actividad laboral y edad mayor de 18 años); (c) son fácilmente 'corregibles', si se verifica que son datos erróneos, o imputables, si son datos *missing* a los que se les puede asignar un determinado valor.

Hay que tener especial cuidado al detectar que la base de datos tiene un porcentaje muy elevado de algún tipo de datos como los anteriores. Si se parte de datos malos, se obtendrán resultados malos, por muy buenos que sean los algoritmos o herramientas que se utilicen en el análisis de datos.

Especial cuidado ha de tenerse con los *outliers*, ya que éstos pueden ser datos extremos en el dominio en el que se mueva la variable en cuestión, pero no por ello han de ser erróneos. Habrá que considerar si se tienen en cuenta en el análisis estadístico o no. Dependiendo de ello, se obtendrán unos resultados u otros, pero ninguno tiene por qué ser erróneo.

Por ejemplo, si se estudia la tendencia del valor medio de los inmuebles en España en una determinada zona desde el año 2000 hasta el año 2021, se comprobará que a partir de 2008 hasta 2013 aparecen cierto valores alejados de los valores habituales. Estos datos corresponden a la crisis económica entre dichos años, por lo que eliminar u obviar

dichos valores nos limitaría la capacidad de predicción a largo plazo, bajo el supuesto posible de la existencia de otro período similar en unos años.

Por tanto, los datos consistentes son aquellos que son correctos y válidos para realizar sobre ellos un análisis estadístico, no existiendo valores *missing*, *outliers*, valores especiales y errores. Si al hacer la exploración de consistencia de nuestros datos detectamos alguno de estos tipos, entonces se procederá a eliminarse o corregirse.

Entendemos por corregirse las diferentes técnicas existentes de transformaciones de datos, como pueden ser las siguientes, las cuales vamos a enumerar pero no vamos a entrar en detalle:

1. Estandarizar variables;
2. Aplicar logaritmos a los datos;
3. Corregir de forma deductiva, al detectar errores de transcripción de datos, por ejemplo.
4. Imputar datos, dependiendo de la información a priori existente o del conocimiento del ámbito de estudio:
   (a) Básico: media, razón (por eejmplo entre promedios de las variables X e Y), regresión lineal (partiendo de los coeficientes estimados β_j para variables auxiliares Y_j);
   (b) Duplicación, tal que se copian valores ya existente a los valores perdidos, siendo el valor estimado igual al valor observado; pudiéndose realizar esta técnica de forma aleatoria, secuencial o por media predictiva;
   (c) Vecino más cercano (KNN del inglés *K nearest neighbour*), utilizando ciertas distancias (euclídea, Gower, Hanning) para cuantificar dicha cercanía;
   (d) Ajuste del valor mínimo, en el que una vez realizada alguna de las anteriores técnicas de imputación se vuelven a ajustar los datos aplicando una distancia euclídea ponderada.

En cuanto a la exploración de datos propiamente dicha, una vez realizada la fase previa de depuración y preprocesamiento de los datos de partida que nos han asegurado la consistencia de los mismos, se lleva a cabo mediante técnicas estadísticas típicamente descriptivas, que permitan analizar de forma gráfica y numérica, entre otras cosas:

- distribución de los datos en relación a cada tipo de variable, a nivel de representación gráfica: histogramas, nube de puntos, cajas y bigotes;
- medidas de posición central (media, mediana, moda), medidas de posición no central (percentiles, cuartiles);
- medidas de dispersión (recorrido, recorrido intercuartílico, varianza, desviación típica, coeficiente de variación de Pearson);
- medidas de forma (simetría de Pearson y Fisher, apuntamiento o curtosis);

- existencia o no de normalidad;
- posibles correlaciones existentes.

También se pueden utilizar técnicas de visualización de datos más avanzadas a nivel unidimensional o multidimensional. Entre las herramientas de visualización tenemos:

1. Interactivas: Datawrapper, Tableau, CARTO, PowerBi, Google Fusion Tables y librerías JavasScript D3.js;
2. Datos cualitativos: Atlas-ti;
3. Datos cuantitativos: SIGIL, UCREL;
4. Basadas en redes (modelos gráficos probabilísticos (PGM))
   (a) GEPHI, Pajek, UCInet, NodeXL;
   (b) librerías gráficas de R (network, igraph, tidygraph, ggraph, visNetwork, networkD3, qgraph, sna, tnet, statnet y NetworkX);
   (c) librerías gráficas de Python (Matplotlib, estándar y la más conocida; Seaborn, basada en matplotlib, especializada en la visualización de datos estadísticos; Bokeh; NumPy; Pandas; scikit-learn; TensorFlow; Keras)

En la Figura 27.2 se muestra una exploración visual de datos, multidimensional, en la que se aprecian los agrupamientos entre las distintas variables y las distancias entre diferentes grupos de ellas. Las aristas de los nodos, que son las variables, son las entidades de relación definidas.

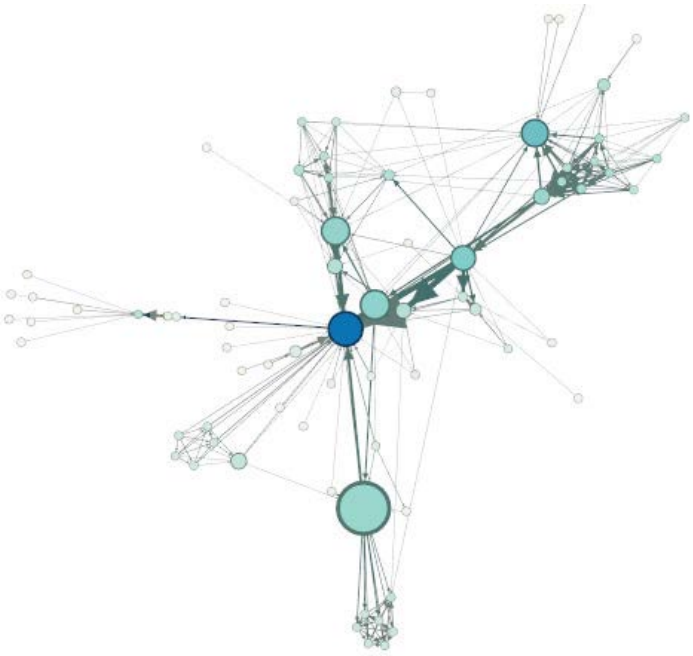

Figura 27.1: Exploración visual de datos (Grafo multidimensional).

.

27.3 Visión básica de estrategias de clasificación, árboles de decisión

27.3.1 Teoría de la decisión estadística

En esta sección se desarrolla una pequeña cantidad de teoría que proporciona un marco para desarrollar modelos que pueden ser conocidos informalmente hasta ahora. En primer lugar, teniendo de referencia el mundo de las variables aleatorias y los espacios de probabilidad, se considera un resultado cuantitativo.

Se sabe que $X \in \mathbb{R}^p$ denota un vector de entrada aleatorio de valor real, e $Y \in \mathbb{R}$ una variable de salida aleatoria de valor real, con distribución conjunta $\mathbb{P}(X, Y)$. Se busca una función $f(X)$ para predecir Y a partir de los valores de entrada dados X. Esta teoría requiere una función de pérdida $L(Y, f(X))$ para penalizar los errores en la predicción y, para ello, la función más común y conveniente es la función de pérdida del error al cuadrado: $L(Y, f(X)) = (Y - f(X))^2$. Esto nos da un criterio para elegir f,

$$\mathrm{EPE}(f) = \mathbb{E}(Y - f(X))^2 \tag{27.1}$$

$$= \int [y - f(x)]^2 \mathbb{P}(dx, dy) \tag{27.2}$$

siendo el error de predicción (al cuadrado) esperado (EPE). Por la condición[1] sobre X, se puede escribir el EPE como

$$\mathrm{EPE}(f) = \mathbb{E}_X \mathbb{E}_{Y|X} \left([Y - f(X)]^2 \mid X \right) \tag{27.3}$$

observando que es suficiente con minimizar el EPE de forma puntual:

$$f(x) = \mathrm{argmin}_c \, \mathbb{E}_{Y|X} \left([Y - c]^2 \mid X = x \right) \tag{27.4}$$

Entonces, la solución es

$$f(x) = \mathbb{E}(Y \mid X = x) \tag{27.5}$$

la esperanza condicionada, también conocida como función de regresión. Por tanto, la mejor predicción de Y en cualquier punto $X = x$ es la media condicionada, cuando la mejor es medida según el error cuadrático medio (ECM).

El método del vecino más cercano intenta implementar directamente esta 'receta' utilizando los datos de entrenamiento. Para cada pundo x, se puede tener el promedio de

[1] se factoriza la densidad conjunta $\Pr(X, Y) = \Pr(Y \mid X) \Pr(X)$ donde $\Pr(Y \mid X) = \Pr(Y, X)/\Pr(X)$, y se divide la integral bivariable en consecuencia.

todos los y_i con valor de entrada $x_i = x$. Dado que normalmente hay como máximo una observación en cualquier punto x, tendríamos

$$\widehat{f}(x) = \text{Ave}\left(y_i \mid x_i \in N_k(x)\right) \tag{27.6}$$

donde el operador 'Ave' denota el valor medio (del inglés *average*), y $N_k(x)$ es el vecino que contiene los k puntos en T más cercanos a x. Dos aproximaciones se están dando aquí:

- el valor esperado es aproximadamente la media de los datos muestrales;
- la condición puntual se relaja a la condición de alguna región 'cercana' al punto objetivo.

Para un tamaño muestral de entrenamiento N grande, es probable que los puntos vecinos estén cerca de x, y a medida que k es mayor, el promedio se estabilizará. De hecho, en condiciones de poca regularidad en la distribución de probabilidad conjunta $\mathbb{P}(X, Y)$, se puede observar que $N, k \to \infty$ es tal que $\frac{k}{N} \to 0$, $\widehat{f}(x) \to \mathbb{E}(Y \mid X = x)$.

A la luz de esto, ¿por qué no mirar más allá, ya que parece que tenemos una aproximación universal? Debido a que a menudo no tenemos muestras muy grandes para ser utilizadas. Entonces, si el modelo lineal o algún modelo más estructurado es apropiado, normalmente se puede obtener una estimación más estable que los vecinos más cercanos de k, aunque ese conocimiento también debe aprenderse de los datos disponibles. Sin embargo, existen otros problemas, a veces mucho peores. En el caso de tener un escenario de datos de altas dimensiones, se observa que a medida que aumenta la dimensión de p, aumenta también la métrica del vecino más cercano k, lo cual parece lógico. Así que podríamos usar el vecino más cercano en caso de que la condición inicialmente planteada nos falle. De esta forma, la convergencia anterior se mantendría, aunque teniendo en cuenta que la tasa de convergencia disminuye a medida que aumenta la dimensión.

¿Cómo encaja la regresión lineal en este entorno? La explicación más simple es que se supone que la función de regresión $f(x)$ es aproximadamente lineal en relación a sus argumentos:

$$f(x) \approx x^T \beta \tag{27.7}$$

Éste es un enfoque basado en modelos en el que se especifica un modelo para la función de regresión. Conectando este modelo lineal para $f(x)$ en el EPE 27.1 y diferenciando, teóricamente se puede resolver para β:

$$\beta = \left[\mathbb{E}\left(XX^T\right)\right]^{-1} \mathbb{E}(XY) \tag{27.8}$$

Hay que tener en cuenta que no hemos condicionado a X; más bien hemos utilizado nuestro conocimiento de la relación funcional para agrupar los valores de X. La solución

de mínimos cuadrados equivale a reemplazar la esperanza en 27.8 por valores medios en los datos de entrenamiento.

Entonces, tanto los k vecinos más cercanos como los mínimos cuadrados terminan aproximándose a las expectativas condicionales mediante promedios. Pero difieren totalmente en cuanto a las asunciones del modelo:

- Mínimos cuadrados asume que $f(x)$ está bien aproximada por una función globalmente lineal.
- k vecinos más cercanos asume que $f(x)$ está bien aproximada por una función localmente constante.

Aunque el último parece más aceptable, hemos visto que podemos pagar un precio por esta flexibilidad.

Muchas de las técnicas más modernas se basan en modelos, siendo mucho más flexibles que el modelo lineal. Por ejemplo, los modelos aditivos asumen que:

$$f(X) = \sum_{j=1}^{p} f_j\left(X_j\right) \tag{27.9}$$

Esto retiene la aditividad del modelo lineal, si bien, cada función de coordenadas f_j es arbitraria. Entonces, la estimación óptima para el modelo aditivo utiliza técnicas como k vecinos más cercanos para aproximar las expectativas condicionadas $univariables$ simultaneamente para cada función de coordenadas. Por lo tanto, los problemas de estimar una expectativa condicional en entornos de altas dimensiones altas se eliminan, en este caso, al imponer algunas hipótesis del modelo (a menudo poco realistas), en este caso la aditividad.

¿Nos parece bien el criterio 27.3? ¿Qué pasa si reemplazamos la función de pérdida L_2 por la función $L_1 : E|Y - f(X)|$? La solución en ese caso es la mediana condicional,

$$\widehat{f}(x) = \text{mediana}(Y \mid X = x) \tag{27.10}$$

que es una medida diferente de ubicación, y sus estimaciones son más robustas que las de la media condicional. El criterio L_1 tiene discontinuidades en sus derivadas, lo que ha dificultado su uso generalizado. Existen otras funciones de pérdida más resistentes, pero el error cuadrático es apropiado analíticamente y el más popular.

¿Qué hacemos cuando el resultado es una variable categórica G? Pues el mismo paradigma funciona aquí, excepto que necesitamos una función de pérdida diferente para penalizar los errores de predicción. Una estimación $\widehat{G}$ asumirá valores en $\mathcal{G}$, el conjunto de clases posibles. Nuestra función de pérdida puede ser representada por una matriz $\mathbf{L}$ de dimensión $K \times K$, donde $K = \text{card}(\mathcal{G})$. $\mathbf{L}$ será cero en la diagonal y no negativa en el resto, donde $L(k, \ell)$ es el 'precio pagado' por clasificar una observación

que pertenece a la clase $\mathcal{G}_k$ como $\mathcal{G}_\ell$. La mayoría de las veces usamos la función de pérdida cero-uno, donde todas las clasificaciones erróneas se cargan en una sola unidad. El error de predicción esperado es,

$$\text{EPE} = \mathbb{E}[L(G, \widehat{G}(X))] \tag{27.11}$$

donde otra vez, la esperanza se toma con respecto a la distribución conjunta $\Pr(G, X)$. Nuevamente tomamos la condición, y podemos escribir el EPE como

$$\text{EPE} = \mathbb{E}_X \sum_{k=1}^{K} L\left[\mathcal{G}_k, \widehat{G}(X)\right] \mathbb{P}\left(\mathcal{G}_k \mid X\right) \tag{27.12}$$

En la Figura 27.2 se observa el límite de decisión óptimo bayesiano para una distribución de datos simulados supuestos, clasificados en círculos rojos y azul. Suponiendo conocida la densidad de generación para cada clase, este límite se puede calcular exactamente.

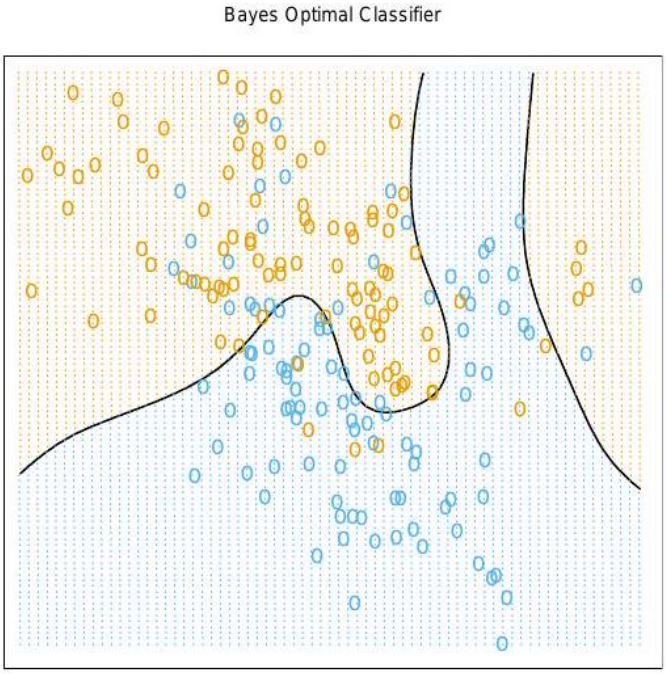

Figura 27.2: Límite del clasificador óptimo bayesiano.

Por tanto, a partir de 27.12 es suficiente minimizar nuevamente el EPE de forma puntual:

$$\widehat{G}(x) = \text{argmin}_{g \in \mathcal{G}} \sum_{k=1}^{K} L\left(\mathcal{G}_k, g\right) \mathbb{P}\left(\mathcal{G}_k \mid X = x\right) \tag{27.13}$$

Con la función de pérdida $0-1$ esto se simplifica a

$$\widehat{G}(x) = \text{argmin}_{g\in\mathcal{G}}[1 - \mathbb{P}(g \mid X = x)] \tag{27.14}$$

o simplemente

$$\widehat{G}(X) = \mathcal{G}_k \text{ si } \mathbb{P}\left(\mathcal{G}_k \mid X = x\right) = \max_{g\in\mathcal{G}} \mathbb{P}(g \mid X = x) \tag{27.15}$$

Esta solución razonable se conoce como clasificador bayesiano, según el cual podemos clasificar a la clase más probable, usando la distribución condicional (discreta) $\mathbb{P}(G \mid X)$. La tasa de error del clasificador bayesiano se denomina *tasa Bayes*.

De nuevo vemos que el clasificador del k vecino más cercano aproxima directamente a la solución más votada en el vecino más cercano, lo que equivale exactamente a esto. Excepto que, la probabilidad condicionada en un punto se relaja a la probabilidad condicionada dentro de un conjunto de vecinos, en un punto, y las probabilidades se estiman mediante proporciones muestrales de entrenamiento.

Supongamos que para un problema de dos clases hemos adoptado el enfoque de variable ficticia *dummy* y hemos codificado G de forma binaria Y, seguida de la estimación de pérdida de error al cuadrado. Entonces, $\widehat{f}(X) = \mathbb{E}(Y \mid X) = \mathbb{P}\left(G = \mathcal{G}_1 \mid X\right)$ si $\mathcal{G}_1$ correspondía a $Y = 1$. Lo mismo ocurre para un problema de clase K, $\mathbb{E}\left(Y_k \mid X\right) = \mathbb{P}\left(G = \mathcal{G}_k \mid X\right)$. Esto muestra que nuestro procedimiento de regresión de variables *dummy*, seguido de la clasificación ajustada al valor más votado, es otra forma de representar el clasificador Bayes. Aunque esta teoría es exacta, en la práctica pueden surgir problemas, dependiendo del modelo de regresión utilizado

Por ejemplo, cuando se usa regresión lineal, es necesario que $\widehat{f}(X)$ sea no positivo, y podríamos ser escépticos en usarlo como la estimación de una probabilidad. En definitiva, existe una gran variedad de enfoques para modelar $\mathbb{P}(G \mid X)$.

27.3.2 Árboles de decisión

Los métodos basados en árboles consisten en segmentar el espacio de predictores en varias regiones. Dentro de cada región, se utiliza la media o la moda de las observaciones de entrenamiento para hacer la predicción. Se dice que son 'métodos basados en árboles' porque las reglas que se utilizan para dividir el espacio de predictores pueden ser representadas en forma de diagrama de árbol. El método más sencillo es el árbol de decisión básico. Luego existen otros métodos como $bagging$, $random\ forest$ y $boosting$ que están basados en el árbol básico pero que mejoran la precisión de este modelo.

En la Figura 27.3 se aprecia un árbol básico, tanto en formato de regiones sobre el plano X-Y, como en subdivisiones lineales de arriba hacia abajo.

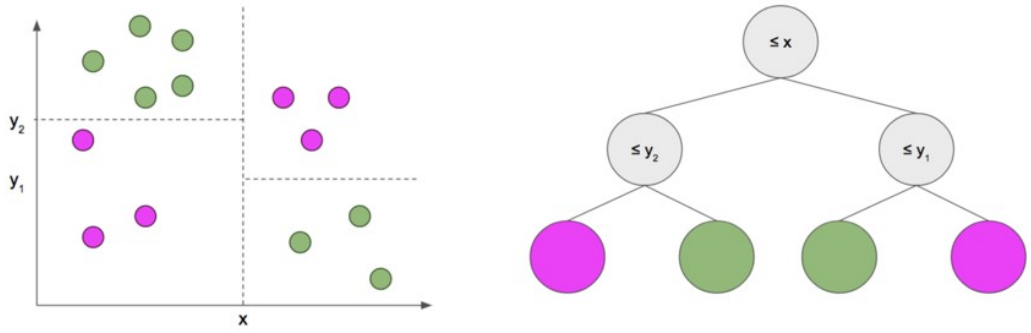

Figura 27.3: Árbol básico.

Otro ejemplo puede verse en la Figura 27.4 en el que se ha aumentado el número de reglas de división. Al igual que antes, está graficado de ambas formas.

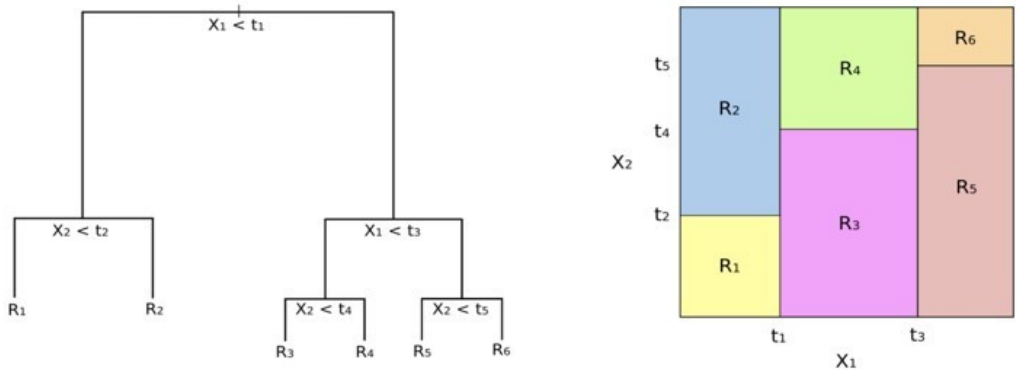

Figura 27.4: Árbol de decisión con 6 divisiones.

Los métodos de árboles se pueden utilizar tanto en clasificación como en regresión. Llegará un momento en que el modelo es tan flexible que comience a seguir las señales de ruido, variabilidad inexplicable dentro de una muestra de datos, y esto dará lugar a un sobreajuste del modelo (*overfitting*). Los métodos basados en árboles son simples y útiles para la interpretación, sin embargo, normalmente no son competitivos comparados con los mejores enfoques de aprendizaje supervisado, como los Modelos lineales de selección y regularización en términos de precisión de predicción, como son:

- Selección por subconjuntos (*Subset Selection*);
- Métodos de contracción (*Shrinkage Methods*);
- Métodos de reducción de dimensión (*Dimension Reduction Methods*);
- Métodos para Altas Dimensiones (*High-Dimensional Methods*);
- Métodos no lineales.

En cuanto a los elementos básicos de un árbol, de decisión y/o clasificación, son los siguientes:

- Nodo raíz: población completa o muestra;

- Ramificación;
- Nodo de decisión;
- Nodo terminal y hoja;
- Poda;
- Rama/sub-árbol;
- Nodos padre e hijo;

El enfoque de árbol de clasificación y regresión (CART del inglés *classification and regression tree*) fue desarrollado por Breiman e Ihaka 1984 y son un tipo de algoritmos de aprendizaje supervisado (por ejemplo, existe una variable objetivo predefinida). Se usan principalmente en problemas de clasificación. Las variables de entrada y salida pueden ser categóricas o continuas, y dividen el espacio de predictores (variables independientes) en regiones distintas y no superpuestas. Es decir, se divide la población N o la muestra n en conjuntos homogéneos basados en la variable de entrada más significativa. La construcción del árbol sigue un enfoque de división binaria recursiva, ó aproximación 'greedy' de arriba-abajo. Aquí hay que diferenciar ambas estrategias de división:

- *Top Down*, se refiere a algoritmos de aprendizaje automático que analizan el conjunto de datos completo en lugar de subconjuntos, a diferencia de, por ejemplo, las técnicas de agrupación;
- *Greedy*, se refiere a que sólo se mira una etapa hacia adelante, sin intenta optimizar globalmente, es decir, analiza la mejor variable para ramificación sólo en el proceso de división actual.

Ambas estrategias de división funcionan particularmente bien juntas, pero no necesariamente tienen por qué. Todos los algoritmos principales del árbol de decisiones son esencialmente una combinación de arriba hacia abajo: se mira todo el conjunto de datos y se va dividiendo en dos; y *greedy*, donde sólo se mira el árbol de decisión un nodo a la vez.

En cuanto al proceso de **construcción de un árbol básico** de regresión, básicamente se siguen dos pasos:

1. Dividimos el espacio de predictores, esto es, el grupo de los valores posible de $X_1, ..., X_p$ en J regiones distintas y que no se solapen, y las llamamos $R_1, ..., R_j$;
2. Para cada observación que caiga dentro de R_j haremos la misma predicción, que será la media de la respuesta de todas las observaciones que caen en la misma región R_j.

Por ejemplo, supongamos que en el Paso 1 obtenemos dos regiones, R_1 y R_2, y que la respuesta media de las observaciones de entrenamiento en la primera región es 10, mientras que la respuesta la media de las observaciones de entrenamiento en la segunda región es 20. Entonces, para una observación dada $X = x$, si $x \in R_1$ predeciremos un valor de 10, y si $x \in R_2$ predeciremos un valor de 20.

Ahora elaboramos el Paso 1 anterior. ¿Cómo construimos las regiones $R_1, \ldots, R_J$? En teoría, las regiones podrían tener cualquier forma. Sin embargo, elegimos dividir el espacio del predictor en rectángulos o cajas de alta dimensión, por simplicidad y facilidad de interpretación del modelo predictivo resultante. El objetivo es encontrar cajas $R_1, \ldots, R_J$ que minimicen la suma de los cuadrados de los residuos (RSS del inglés *Residual Square Sum*), dada por

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \widehat{y}_{R_j}\right)^2 \tag{27.16}$$

donde $\widehat{y}_{R_j}$ es la respuesta media para las observaciones de entrenamiento dentro de la j-ésima caja (región o partición). Desafortunadamente, es computacionalmente inviable considerar cada posible partición del espacio de características en J cajas. Por esta razón, adoptamos una aproximación '*greedy*' de arriba-abajo que se conoce como *división binaria recursiva*. El enfoque es de arriba hacia abajo porque comienza en la parte superior del árbol, en cuyo punto binario todas las observaciones pertenecen a una sola región, y luego divide sucesivamente el espacio de predicción; cada división se indica mediante dos nuevas ramas más abajo en el árbol. Es *greedy* porque en cada paso del proceso de construcción del árbol, la mejor división se realiza en ese paso en particular, en lugar de mirar hacia adelante y elegir una división que conducirá a un mejor árbol en algún paso futuro.

Para realizar una división binaria recursiva, primero seleccionamos el predictor X_j y el punto de corte s tal que dividir el espacio del predictor en las regiones $\{X \mid X_j < s\}$ [2] y $\{X \mid X_j \geq s\}$ conduce a la mayor reducción posible del RSS. Es decir, consideramos todos los predictores $X_1, \ldots, X_p$, y todos los posibles valores del punto de corte s para cada uno de los predictores, y luego elegimos el predictor y el punto de corte de manera que el árbol resultante tiene el RSS más bajo. Con mayor detalle, para cualquier j y s, definimos el par de semiplanos

$$R_1(j, s) = \{X \mid X_j < s\} \text{ y } R_2(j, s) = \{X \mid X_j \geq s\} \tag{27.17}$$

y buscamos el valor de j y s que minimiza la ecuación

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \widehat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \widehat{y}_{R_2})^2 \tag{27.18}$$

donde $\widehat{y}_{R_1}$ es la respuesta media para las observaciones de entrenamiento en $R_1(j, s)$, e $\widehat{y}_{R_2}$ es la respuesta media para las observaciones de entrenamiento en $R_2(j, s)$. Encontrar los valores de j y s que minimizan 27.18 puede hacerse bastante rápido, especialmente cuando el número de características p no es demasiado grande.

[2] La notación $\{X \mid X_j < s\}$ significa la *región del espacio del predictor en la que X_j toma un valor menor que s.*

Después, repetimos el proceso, buscando el mejor predictor y el mejor punto de corte para dividir aún más los datos y minimizar el RSS dentro de cada una de las regiones resultantes. Sin embargo, esta vez, en lugar de dividir todo el espacio del predictor, dividimos una de las dos regiones previamente identificadas. Ahora tenemos tres regiones. Nuevamente, buscamos dividir aún más una de estas tres regiones, para minimizar el RSS. El proceso continúa hasta que se alcanza un criterio de parada; por ejemplo, podemos continuar hasta que ninguna región contenga más de cinco observaciones.

Una vez que se han creado las regiones $R_1, \ldots, R_J$, predecimos la respuesta para una observación de prueba dada, usando la media de las observaciones de entrenamiento en la región a la que pertenece esa observación de prueba.

En la Figura 27.5, hay un ejemplo de este enfoque, con cinco regiones.

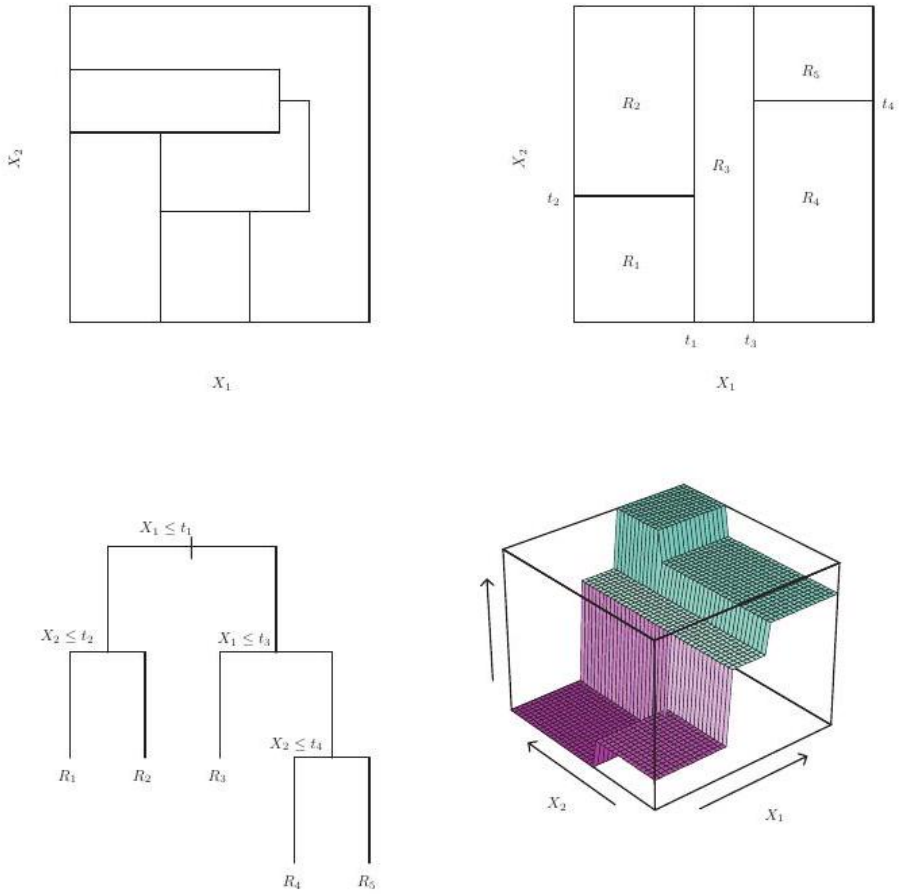

Figura 27.5: **Arriba izquierda**: partición del espacio de características bidimensional que no podría resultar de una división binaria recursiva. **Arriba derecha**: salida de la división binaria recursiva en un ejemplo bidimensional. **Abajo izquierda**: árbol correspondiente a la partición en el panel superior derecho. **Abajo derecha**: gráfica en perspectiva de la superficie de predicción correspondiente a ese árbol.

Poda de árbol

El proceso descrito anteriormente puede producir buenas predicciones en el conjunto

de entrenamiento, pero es probable que sobreajuste los datos, lo que conducirá a un rendimiento deficiente del conjunto de prueba. Esto se debe a que el árbol resultante puede ser demasiado complejo. Un árbol más pequeño con menos divisiones (es decir, menos regiones $R_1, \ldots, R_J$) podría conducir a una menor varianza y una mejor interpretación a costa de un pequeño sesgo. Una posible alternativa al proceso descrito anteriormente es construir el árbol solo mientras la disminución en el RSS debido a cada división exceda algún umbral (alto). Esta estrategia dará como resultado árboles más pequeños, pero es demasiado 'corta', ya que una división aparentemente sin valor al principio del árbol podría ir seguida de una división muy buena, es decir, una división que lleve a una gran reducción del RSS más adelante.

Por lo tanto, una mejor estrategia es hacer crecer un árbol muy grande T_0, y luego podarlo para obtener un subárbol. ¿Cómo determinamos la mejor forma de podar el árbol? Intuitivamente, nuestro objetivo es seleccionar un subárbol que conduzca a la tasa de error de prueba más baja. Dado un subárbol, podemos estimar su error de prueba utilizando la validación cruzada o la validación por aproximación conjunta. Sin embargo, estimar el error de validación cruzada para cada posible subárbol sería demasiado engorroso, ya que existe un número extremadamente grande de posibles subárboles. En cambio, necesitamos una forma de seleccionar un conjunto pequeño conjunto de subárboles para su consideración.

Reducción de la complejidad de costes: también conocida como poda de los eslabones más débiles, nos brinda una manera de hacer precisamente esto en los costes. En lugar de considerar todos los subárboles posibles, consideramos una secuencia de árboles indexada por un parámetro de ajuste no negativo α.

Definición 5

Algoritmo 1. Construcción de un Árbol de Regresión

1. Utilizar la división binaria recursiva para hacer crecer un árbol grande con los datos de entrenamiento, deteniéndose sólo cuando cada nodo terminal tenga menos de un número mínimo de observaciones.
2. Aplicar la poda de coste complejo al árbol grande para obtener una secuencia de los mejores subárboles, en función de α.
3. Usar validación cruzada de K-fold para elegir α. Esto es, dividir los datos de entrenamiento en K grupos. Para cada $k = 1, \ldots, K$:
   (a) Repetimos los Pasos 1 y 2 sobre todos los datos esperados en el k grupo de los datos de entrenamiento.
   (b) Evaluar el error de predicción cuadrático medio en los datos, dejando fuera al k-ésimo grupo, como una función de α

Promediar los resultados para cada valor de α, y escoger α para minimizar el error medio.

4. Devolver el subárbol del paso 2 que corresponde al valor elegido de α

A cada valor de α le corresponde un subárbol $T \subset T_0$ tal que

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \widehat{y}_{R_m})^2 + \alpha |T| \tag{27.19}$$

es lo más pequeño posible. Aquí $|T|$ indica el número de nodos terminales del árbol T, R_m es el rectángulo (es decir, el subconjunto del espacio predictor) correspondiente al m-ésimo nodo terminal, e $\widehat{y}_{R_m}$ es la respuesta predicha asociada con R_m- es decir, la media de las observaciones de entrenamiento en R_m. El parámetro de ajuste α controla una compensación entre la complejidad del subárbol y su ajuste a los datos de entrenamiento. Cuando $\alpha = 0$, entonces el subárbol T simplemente será igual a T_0, porque entonces 27.19 sólo mide el error de entrenamiento. Sin embargo, a medida que α crece, hay un precio que pagar por tener un árbol con muchos nodos terminales, por lo que la cantidad 27.19 tenderá a minimizarse para un subárbol más pequeño. La ecuación 27.19 recuerda a la regresión de Lasso, en la que se utiliza una formulación similar para controlar la complejidad de un modelo lineal.

Resulta que a medida que aumentamos α desde cero en 27.19, las ramas del árbol quedan podadas de forma anidada y predecible, por lo que es fácil obtener la secuencia completa de subárboles en función de α. Podemos seleccionar el valor de α usando un conjunto de datos de validación o usando validación cruzada. Después, volvemos al conjunto de datos completo y obtenemos el subárbol correspondiente a α. Este proceso se resume en el anterior **Algoritmo 5**.

27.4 Conceptos básicos de análisis clúster

Por ejemplo, supongamos que tenemos un conjunto de n observaciones, cada una con p características. Las n observaciones podrían corresponder a muestras de tejido para pacientes con cáncer de mama, y las p características podrían corresponder a las mediciones recogidas para cada muestra de tejido; pueden ser medidas clínicas, como el estadio o el grado del tumor, o pueden ser medidas de expresión génica. Es posible que tengamos una razón para creer que existe cierta heterogeneidad entre las n muestras de tejido; por ejemplo, tal vez haya algunos subtipos desconocidos diferentes de cáncer de mama. La agrupación podría utilizarse para encontrar estos subgrupos.

Se trata de un método no supervisado ya que el proceso ignora la variable respuesta que indica a que grupo pertenece realmente cada observación, si es que existe tal variable. Por el contrario, el objetivo en los problemas supervisados es intentar predecir algún vector de resultado, como el tiempo de supervivencia o la respuesta al tratamiento

farmacológico.

Esta característica diferencia al clustering de las técnicas estadísticas conocidas como análisis discriminante, que emplean un conjunto de entrenamiento en el que se conoce la verdadera clasificación.

Otro algoritmo de aprendizaje no supervisado es el de componentes principales (PCA), y al igual que el clustering, ambos buscar simplificar los datos a través de un pequeño número agrupaciones, si bien, sus mecanismos son diferentes:

- PCA busca encontrar una representación de baja dimensión de las observaciones que explique una buena parte de la varianza;
- Clustering busca encontrar subgrupos homogéneos entre las observaciones.

Respecto a las aplicaciones del clustering, una importante además de la investigación en el ámbito de la salud, se encuentra en el marketing y en los estudios de mercado. En ellos se tiene acceso a un gran número de mediciones y variables (por ejemplo, ingresos medios por hogar, empleo, distancia al área urbana más próxima, etc.) de un gran número de personas. El objetivo en estos estudios es el de construir una *segmentación de mercado* mediante la identificación de subgrupos de personas que pueden ser más receptivos a una campaña publicitaria en concreto, o más motivados a la compra de un producto o servicio. En ese sentido, existen diferentes áreas de mercado en la que se aplican estas técnicas: financiero, automoción, sanidad privada, seguros, telecomunicaciones, transporte,...

En cuanto a las tipologías más representativas del clustering se pueden identificar las siguientes:

- Clúster por particiones: Es necesario que el analista de datos marque inicialmente el número de clústers que se van a construir (K-medias , K-medoides, CLARA (**C**lustering **L**arge **A**pplications)).
- clúster jerárquico: No es necesario que el analista de datos marque inicialmente el número de clústers que se van a construir. La representación final se denomina dendograma, en el que se observa una estructura de árbol en su totalidad, alcanzando un elevado nivel de detalle para cada uno de los conglomerados alcanzados, de 1 a n.
- Métodos mixtos, que combinan o modifican los anteriores (jerárquico, difuso, basado en modelos, basado en densidades).

Del anterior listado en este apartado nos centraremos en dos de las aproximaciones más conocidas de clustering: (1) K-medias, que será desarrollado posteriormente; (2) clúster jerárquico.

En ambos casos surgen una pregunta y una dificultad inicial:

- K-medias, ¿cómo puedo saber a priori el número de grupos o clústers? Este método (MacQueen y col. 1967) forma los diferentes grupos tal que su varianza

interna sea lo más pequeña posible y, por tanto, se minimice la varianza total de todas las observaciones. Para ello, es necesario definir la medida de dicha varianza entre los clústers.

- Jerárquico, si desconozco el número de clústers que quiero o necesito, ¿hasta qué nivel de detalle puedo asumir en el número de ramificaciones de mi dendograma para ser capaz de realizar una interpretación de mis datos?

Evidentemente, hay ventajas y desventajas en cada uno de estos enfoques de agrupamiento. En general, se pueden agrupar observaciones sobre la base de las características para identificar subgrupos entre las observaciones, o se pueden agrupar características sobre la base de las observaciones para descubrir subgrupos entre las características. Por simplicidad, nos centraremos en la agrupación de observaciones sobre la base de las características, aunque se puede realizar lo contrario, simplemente transponiendo la matriz de datos. Por ese motivo, a continuación se desarrolla el algoritmo de las K-medias.

27.4.1 K-medias

El método de las K-medias es una aproximación sencilla para dividir los datos de partida en k grupos distintos, sin solapar grupos. Para realizar la agrupación de las K-medias, primero debemos especificar el número deseado de agrupaciones K; entonces el algoritmo asignará cada observación exactamente a uno de los K grupos. En la Figura 27.6, hay un ejemplo de esta construcción de las K-medias con una simulación de 150 observaciones en dos dimensiones, usando tres valores diferentes para K (k=2, k=3 y k=4).

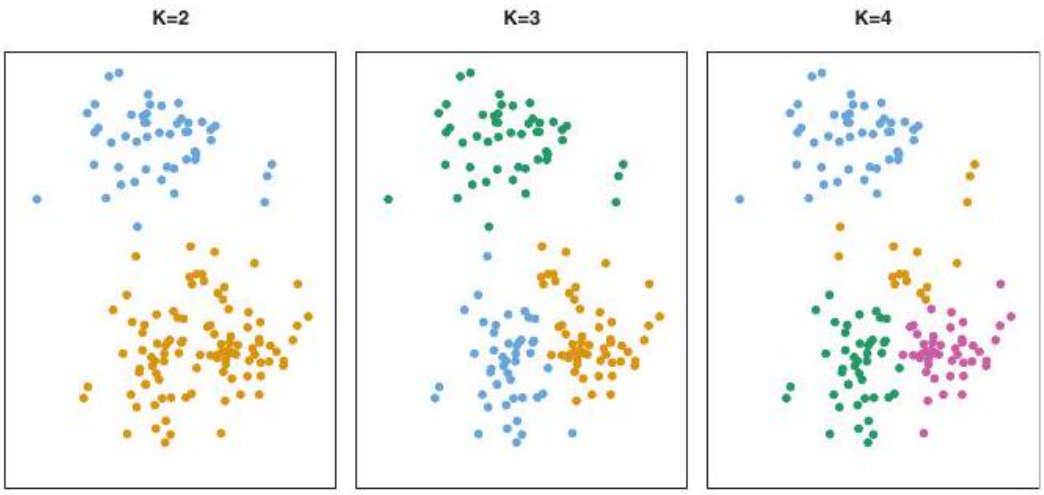

Figura 27.6: Conjunto de 150 observaciones simuladas en un espacio bidimensional. Los paneles muestran los resultados de aplicar K-medias con diferentes valores de k, el número de clústers. El color de cada observación indica el grupo al que se asignó mediante el algoritmo de K-medias. Téngase en cuenta que no hay ningún orden de clústeres, por lo que el color del clúster es arbitrario.

El procedimiento de las K-medias proviene de un problema matemático simple e intuitivo. Empezamos definiendo alguna notación. Sean $C_1, \ldots, C_K$ que denotan conjuntos que contienen los índices de las observaciones en cada grupo. Estos conjuntos satisfacen dos propiedades:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. En otras palabras, cada observación pertenece al menos a uno de los K grupos.

2. $C_k \cap C_{k'} = \emptyset$ para todo $k \neq k'$. Es decir, los grupos no se superponen: ninguna observación pertenece a más de un grupo.

Por ejemplo, si la observación i-ésima en el grupo k-ésimo, entonces $i \in C_k$. La idea que subyace es que el método de las K-medias significa que una buena agrupación es aquella en la que la variación dentro del grupo (intra-cluster) es lo más pequeña posible.

La varianza intra-grupo para el grupo C_k es una medida $W(C_k)$ de la cantidad en que las observaciones dentro de un conglomerado difieren entre sí. Por eso queremos resolver el problema de minimización, tal que:

$$\underset{C_1,\ldots,C_K}{\text{minimizar}} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \tag{27.20}$$

En otras palabras, esta fórmula dice que queremos dividir las observaciones en K grupos tal que la varianza total entre grupos, sumando todos los K grupos, sea la menor posible. Que es la idea de partida de este apartado

Resolver 27.20 parece una idea razonable, pero para que sea factible, se necesita definir la variación intra-grupo. Hay muchas formas posibles de definir este concepto, pero, con mucho, la opción más común implica la distancia euclídea al cuadrado. Es decir, se define

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{27.21}$$

donde $|C_k|$ denota el número de observaciones en el k-ésimo grupo. En otras palabras, la varianza intra-grupos para el k-ésimo grupo es la suma de es la suma de todas las distancias euclídeas cuadráticas, por pares, entre las observaciones en el k-ésimo grupo, dividido por el número total de observaciones en el k-ésimo grupo. Combinando 27.20 y 27.21 se tiene el problema de optimización que define el método de las K-medias,

$$\underset{C_1,\ldots,C_K}{\text{minimizar}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \tag{27.22}$$

Ahora, nos gustaría encontrar un algoritmo para resolver 27.22, es decir, un método para dividir las observaciones en K grupos tal que el objetivo de 27.22 sea minimizado.

Si bien, esto es un problema muy difícil de solucionar con precisión, ya que hay casi K^n formas de dividir n observaciones en K grupos. Y éste es un número gigante a menos que K y n sean pequeños. Afortunadamente, se puede demostrar que un algoritmo muy simple proporciona un óptimo local, una solución bastante buena, al problema de optimización de las K-medias en 27.22. Esta aproximación se presenta en el **Algoritmo 6**.

Definición 6

Algoritmo 2. Cluster K-medias

- Paso 1. Asignar de forma aleatoria un número, entre 1 y K, a cada una de las observaciones. Estos números sirven de asignación inicial de grupos a las observaciones;
- Paso 2. Iterar hasta que las asignaciones de clústeres dejen de cambiar:
  - (a) Para cada grupo K, se calcula el centroide, punto único de representación, del grupo. El centroide del k-ésimo grupo es el vector de medias de las p características para las observaciones en el k-ésimo grupo;
  - (b) Asignar cada observación al grupo cuyo centroide sea el más cercano, utilizando la distancia euclídea.

Este **Algoritmo 6** garantiza disminuir el valor del objetivo de 27.22 en cada paso. Para entender por qué, la siguiente igualdad lo aclara:

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 \tag{27.23}$$

donde $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ es la media de la característica j en el grupo C_k. En el Paso 2(a) las medias de los grupos para cada característica son contantes que minimizan las desviaciones de la suma de cuadrados, y en el Paso 2(b), reasignar las observaciones solo puede mejorar 27.23. Esto significa que a medida que se ejecuta el algoritmo, la agrupación obtenida mejorará continuamente hasta que el resultado ya no cambie; la función objetivo de 27.22 nunca aumentará. Cuando el resultado ya no cambia, se ha alcanzado un óptimo local.

La Figura 27.7 muestra la progresión del algoritmo en el ejemplo de la Figura 27.6. El nombre del método de las k-medias proviene del echo de que en el Paso 2(a), los centroides de los grupos se calculan como la media de las observaciones asignadas a cada grupo.

Debido a que el algoritmo de las K-medias encuentra un óptimo local en lugar de global, los resultados obtenidos dependerán de la asignación inicial, que es aleatoria, de cada observación a los grupos en el Paso 1 del **Algoritmo 6**. Por esta razón, es importante

ejecutar el algoritmo varias veces desde diferentes configuraciones iniciales.

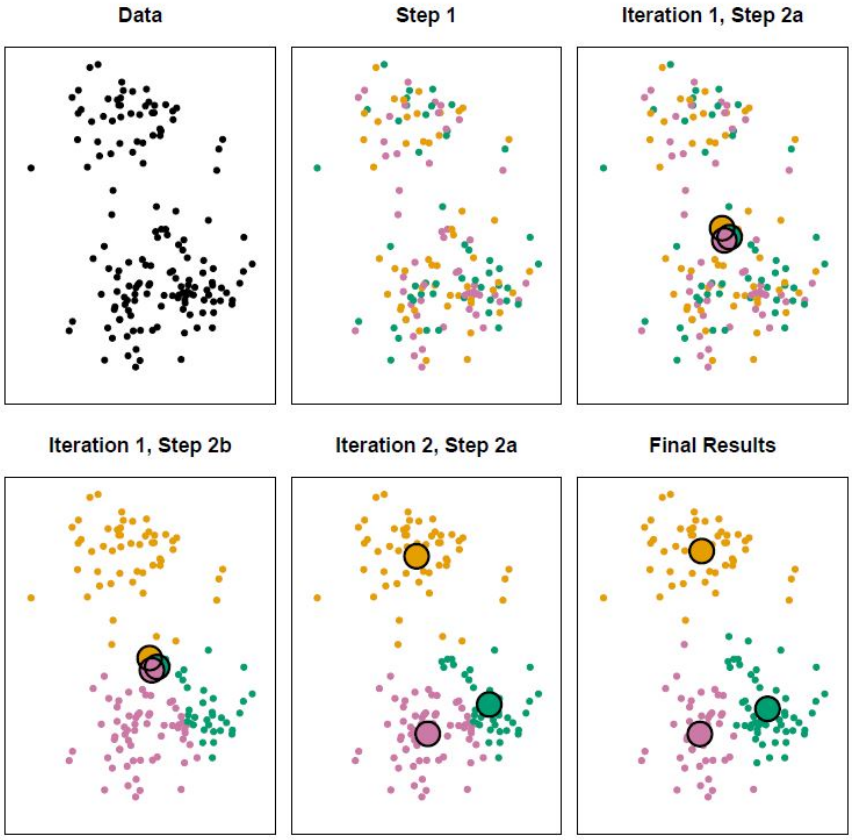

Figura 27.7: Progreso del algoritmo de las K-medias en el ejemplo anterior 27.6 con K=3. **Arriba izquierda**: observaciones. **Arriba centro**: Paso 1 del algoritmo, cada observación se asigna aleatoriamente a un grupo. **Arriba derecha**: Paso 2(a), se calculan los centroides de los grupos. Estos se representan con un gran círculo coloreado. Inicialmente, los centroides se superponen casi por completo porque las asignaciones de conglomerados iniciales se eligieron al azar. **Abajo izquierda**: Paso 2(b), cada observación es asignada al centroide más cercano. **Abajo centro**: Paso 2(a) se realiza una vez más, lo que lleva a nuevos centroides. **Abajo derecha**: resultados obtenidos tras 10 iteraciones.

Luego se selecciona la mejor solución, es decir, aquella para la que la función objetivo de 27.22 es la menor posible. La Figura 27.8 muestra el local óptimo obtenido ejecutando el método de las K-medias 6 veces, usando 6 diferentes asignaciones iniciales de grupos, partiendo de los datos de ejemplo de la Figura 27.6. En ese caso, la mejor agrupación es aquella en la que el valor de la función objetivo es 235,8.

Como se ha visto, para desarrollar el método de las K-medias es necesario decidir cuántos grupos esperamos tener en los datos. Por lo que el problema de seleccionar el número K está lejos de ser sencillo.

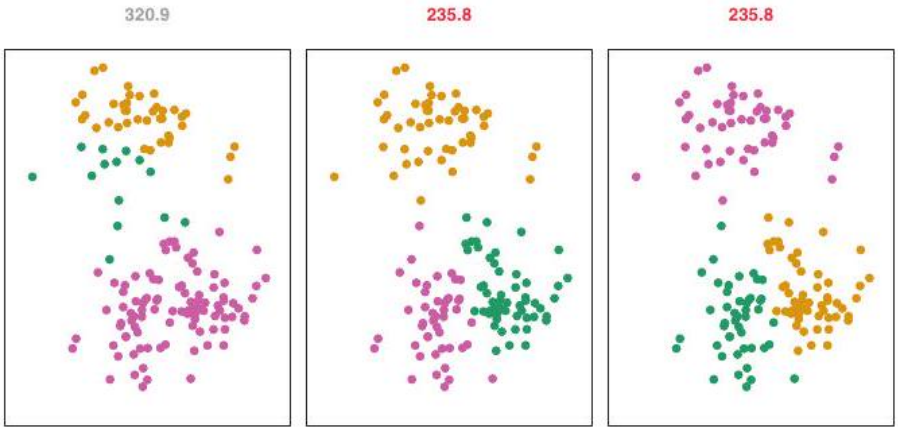

Figura 27.8: Progreso del algoritmo de las K-medias en el ejemplo anterior 27.6 con k=3, cada vez con una asignación aleatoria diferentes de las observaciones en el Paso 1 del algoritmo de las K-medias. Sobre cada gráfico se muestra el valor de la función objetivo de 27.22. Se obtienen 3 óptimos locales diferentes, de los que uno de ellos resulta ser el menor de la función objetivo y proporciona la mejor separación entre los grupos. Todos aquellos etiquetados en rojo alcanzan la misma mejor solución, con un valor de la función objetivo de 235,8.

Bibliografía

Breiman, Leo y Ross Ihaka (1984). *Nonlinear discriminant analysis via scaling and ACE*. Department of Statistics, University of California.

De Jonge, Edwin y Mark Van Der Loo (2013). *An introduction to data cleaning with R*. Statistics Netherlands Heerlen.

Hastie, Trevor, Robert Tibshirani y Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

James, G., D. Witten, T. Hastie y R. Tibshirani (2013). *An introduction to statistical learning*. Vol. 112. Springer.

MacQueen, James y col. (1967). "Some methods for classification and analysis of multivariate observations". En: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA, págs. 281-297.